(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
30 May 2002 (30.05.2002)

PCT

(10) International Publication Number
WO 02/42485 A2

(51) International Patent Classification[7]: C12Q

(21) International Application Number: PCT/US01/43742

(22) International Filing Date: 21 November 2001 (21.11.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/718,295 21 November 2000 (21.11.2000) US

(71) Applicant *(for all designated States except US)*: **AFFYMETRIX, INC.** [US/US]; 3380 Central Expressway, Santa Clara, CA 95051 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **MEI, Rui** [US/US]; 234 Rodonovan Drive, Santa Clara, CA 95051 (US). **WEBSTER, Teresa** [US/US]; 10002 Pescadero Creek Road, Loma Mar, CA 94021 (US).

(74) Agents: **LIEBESCHUETZ, Joe** et al.; Townsend & Townsend & Crew LLP, 2 Embarcadero Center, 8th Floor, San Francisco, CA 94111 (US).

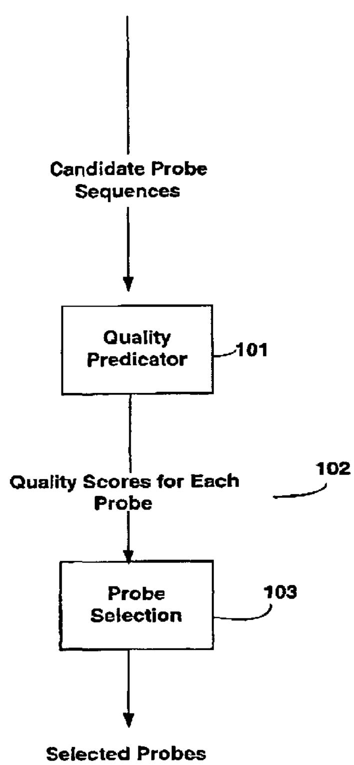
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR,

[Continued on next page]

(54) Title: METHODS AND COMPUTER SOFTWARE PRODUCTS FOR SELECTING NUCLEIC ACID PROBES

(57) Abstract: Methods and computer software products are provided for selecting nucleic acid probes. In one embodiment, perfect match intensity, mismatch intensity and the slope of quantitative response of a probe are predicted. A unified quality score is calculated. Probes are selected based on the unified quality score.



WO 02/42485 A2

GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

— *entirely in electronic form (except for this front page) and available upon request from the International Bureau*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

METHODS AND COMPUTER SOFTWARE PRODUCTS FOR SELECTING NUCLEIC ACID PROBES

RELATED APPLICATIONS

The present application is a continuation of USSN 09/718,295, filed November 21, 2000. Th present application is related to U.S. Patent Application Number 09/721,042, Attorney docket number 3367, entitled "Methods and Computer Software Products for Predicting Nucleic Acid Hybridization Affinity", filed on November 21, 2000, U.S. Patent Application Number 60/252,617, Attorney Docket Number 3373, entitled "Methods and Computer Software Products for Selection Nucleic Acid Probes Using Dynamic Programming", filed on November 21, 2000 and U.S. Patent Application Serial Number 09/745,965, filed on December 21, 2000. The applications are incorporated herein by reference for all purposes.

BACKGROUND OF THE INVENTION

The present invention relates to methods for designing nucleic acid probe arrays. U.S. Patent No. 5,424,186 describes a pioneering technique for, among other things, forming and using high density arrays of molecules such as oligonucleotides, RNA or DNA), peptides, polysaccharides, and other materials. This patent is hereby incorporated by reference for all purposes. There is still great need for methods, systems and software for designing high density nucleic acid probe arrays.

SUMMARY OF THE INVENTION

In one aspect of the invention, computer implemented methods are provided for selecting oligonucleotide probes. The methods including steps of a) predicting hybridization intensities of

a plurality of candidate probes, b) predicting quantitative responses of the candidate probes to the amount of their targets, c) selecting the probes from the candidate probes according to their hybridization intensities and quantitative response and d) spacing the probes along the sequence to avoid overlapping probes.

In some embodiments, the quantitative response is the slope of the response curve of a probe and hybridization intensity (I) is determined using the equation:

$$Ln(I) = \sum_{i=1}^{3N} W_i S_i + C_2$$

Or

$$Ln(I) = \sum_{i=1}^{3N} W_i S_i$$

wherein W_i is a weight coefficient; S_i is a functional of the sequence of a probe; N is the number of bases of a probe; and C_2 is a constant. In some embodiments, the weight coefficient is determined using multiple linear regression analysis.

In some preferred embodiments, the methods for selecting probes further include a step of predicting mismatch hybridization intensities of corresponding mismatch probes of the candidate probes and the selecting step is also based upon the mismatch hybridization intensities. In some cases, the mismatch probes are different from their corresponding candidate probes in one base pair in the middle of their sequences. In preferred embodiments, the match hybridization intensities are predicted according to the sequences of the candidate genes. In some embodiments, mismatch hybridization intensities are determined according to the following equation:

$$Ln(I) = \sum_{i=1}^{3N} W_i^{'} \ S_i + C_2^{'} \quad \text{or}$$

$$Ln(I)=\sum_{i=1}^{3N} W_i^{'}\ S_i$$

wherein said W'_i is a weight coefficient; S_i is a functional of the sequence of the perfect match probe; N is the number of bases of the probe; and $C_2^{'}$ is a constant, and I is the intensity of the mismatch probe.

The method of selecting probes may further include a step of calculating a unified quality score based upon predicted hybridization intensities.

In another aspect of the invention, computer software products are provided for selecting oligonucleotide probes. The software product includes computer program code for predicting hybridization intensities of a plurality of candidate probes; computer program code for predicting quantitative responses of the candidate probes to the amount of their targets; and computer program code for selecting said probes from said candidate probes according to said hybridization intensities and said quantitative response; and a computer readable media for storing said computer program codes.

In some embodiments, the quantitative response is the slope of the response curve of a probe. The hybridization intensity (I) may be determined using the equation:

$$Ln(I)=\sum_{i=1}^{3N} W_i S_i + C_2$$

Or

$$Ln(I)=\sum_{i=1}^{3N} W_i S_i$$

wherein said W_i is a weight coefficient; S_i is a functional of the sequence of a probe; N is the number of bases of a probe; and C_2 is a constant.

The weight coefficient is determined using multiple linear regression analysis.

The computer software product comprising computer program code for predicting mismatch hybridization intensities of corresponding mismatch probes of said candidate probes and wherein said selecting step is also based upon said mismatch hybridization intensities. The method of Claim 13 wherein said mismatch probes are different from their corresponding candidate probes in one base pair in the middle of their sequence. The mismatch hybridization intensities may be predicted according to the sequences of said candidate genes. In some embodiment, the mismatch hybridization intensities are determined according to the following equation:

$$Ln(I) = \sum_{i=1}^{3N} W_i' \; S_i + C_2' \text{ or}$$

$$Ln(I) = \sum_{i=1}^{3N} W_i' \; S_i$$

wherein said W'_i is a weight coefficient; S_i is a functional of said sequence of said probe; N is the number of bases of said probe; and C_2' is a constant. In one additional embodiment, the computer program code for selecting probes include computer program code for calculating a unified score for each probe.

In yet another aspect of the invention, a system for selecting nucleic acid probes is provided. The system include a processor; and a memory being coupled to the processor, the memory storing a plurality machine instructions that cause the processor to perform a plurality of logical steps when implemented by the processor, the logical steps including:

a) predicting hybridization intensities of a plurality of candidate probes;

b) predicting quantitative responses of the candidate probes to the amount of their targets;

c) selecting the probes from the candidate probes according to the hybridization intensities and the quantitative response;

d) spacing the probes along the sequence to avoid overlapping probes.

In some embodiments, the quantitative response is the slope of the response curve of the probe. The hybridization intensity (I) is determined using the equation:

$$Ln(I) = \sum_{i=1}^{3N} W_i S_i + C_2$$

Or

$$Ln(I) = \sum_{i=1}^{3N} W_i S_i$$

wherein said W_i is a weight coefficient; S_i is a functional of the sequence of a probe; N is the number of bases of a probe; and C_2 is a constant. The weight coefficient may be determined using multiple linear regression analysis.

In some preferred embodiments, the logic steps may further include predicting mismatch hybridization intensities of corresponding mismatch probes of the candidate probes and the selecting step is also based upon mismatch hybridization intensities. The mismatch probes are different from their corresponding candidate probes in one base pair in the middle of their sequences. The mismatch hybridization intensities may be predicted according to the sequences of said candidate genes. In some embodiments, the mismatch hybridization intensities are determined according to the following equation:

$$Ln(I) = \sum_{i=1}^{3N} W_i' \ S_i + C_2' \quad \text{or}$$

$$Ln(I) = \sum_{i=1}^{3N} W_i' \ S_i$$

wherein said W'_i is a weight coefficient; S_i is a functional of the sequence of the probe; N is the number of bases of the probe; and C_2' is a constant. The selecting step may also include a step of calculating a unified quality score based upon predicted hybridization intensities.

The present predictive methods are preferably used to select a collection of probes and an array upon which they are used.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings, which are incorporated in and form a part of this specification, illustrate embodiments of the invention and, together with the description, serve to explain the principles of the invention:

FIG. 1 illustrates an example of a computer system that may be utilized to execute the software of an embodiment of the invention.

FIG. 2 illustrates a system block diagram of the computer system of FIG. 1.

FIG. 3 illustrates a model system for probe sequence-based prediction of probe hybridization behavior (probe quality).

FIG. 4 illustrates a basic physical model for probe target interaction.

FIG. 5 shows an example of S_i for an exemplary probe.

FIG.s 6A and 6B show predicted relative ΔG for perfect match and mismatch probes.

FIG. 7 shows an overall reaction of probe target formation.

FIG. 8 shows concentration dependency of hybridization intensity.

FIGs. 9A and 9B show the relationship between probe-target binding affinity (K_{app}) and the slope (S).

FIG. 10 shows an embodiment of a process for selecting probes.

FIG. 11 shows a pool of candidate probes.

FIG. 12 shows another embodiment of a process for selecting probes.

FIG. 13 shows yet another embodiment of a process for selecting probes.

FIG. 14 shows a process for obtaining weight coefficients.

FIG. 15 shows yeast clones used to produce targets.

FIG. 16 shows a Latin Square design.

FIG. 17 shows Latin Square data sets from yeast_test_hyb chips.

FIG. 18 shows a crossvalidation bootstrapping process.

FIGs. 19A, 19B, 20A and 20B show correlation between predicted and observed hybridization intensities for perfect match probes and mismatch probes.

FIG. 21 shows hybridization intensity at different spike concentrations.

FIG. 22 shows correlation between predicted and observed intensities over the entire concentration range.

FIG. 23 shows predicted versus observed intensities for negative control target.

FIG. 24 shows predicted versus observed slopes and improvement of correlation between the two slopes after filtering saturated probes.

FIG. 25 shows prediction of hybridization of a human expression chip to human target sequences using weight coefficients generated from the yeast model system.

FIG. 26 shows distribution of correlation coefficients.

FIG. 27 shows selection of probes using dynamic programming.

FIG. 28 compares different methods for selecting probes.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Reference will now be made in detail to the preferred embodiments of the invention. While the invention will be described in conjunction with the preferred embodiments, it will be understood that they are not intended to limit the invention to these embodiments. On the contrary, the invention is intended to cover alternatives, modifications and equivalents, which may be included within the spirit and scope of the invention.

I. Glossary

"Nucleic acids," according to the present invention, may include any polymer or oligomer of nucleosides or nucleotides (polynucleotides or oligonucletodies), which include pyrimidine and purine bases, preferably cytosine, thymine, and uracil, and adenine and guanine, respectively. See Albert L. Lehninger, PRINCIPLES OF BIOCHEMISTRY, at 793-800 (Worth Pub. 1982) and L. Stryer BIOCHEMISTRY, 4th Ed., (March 1995), both incorporated by reference. Indeed, the present invention contemplates any deoxyribonucleotide, ribonucleotide or peptide nucleic acid component, and any chemical variants thereof, such as methylated, hydroxymethylated or glucosylated forms of these bases, and the like. The polymers or oligomers may be heterogeneous or homogeneous in composition, and may be isolated from naturally-occurring sources or may be artificially or synthetically produced. See U.S. patent application Serial No. 08/630,427 which is incorporated herein by reference in its entirety for all purposes. In addition, the nucleic acids may be DNA or RNA, or a mixture thereof, and may exist permanently or transitionally in single-stranded or double-stranded form, including homoduplex, heteroduplex, and hybrid states. Oligonucleotides and polynucleotides are included in this definition and relate to two or more nucleic acids in a polynucleotide.

"Probe," as used herein, is defined as a nucleic acid, such as an oligonucleotide, capable of binding to a target nucleic acid of complementary sequence through one or more types of chemical bonds, usually through complementary base pairing, usually through hydrogen bond formation. As used herein, a probe may include natural (i.e., A, G, U, C, or T) or modified bases (7-deazaguanosine, inosine, etc.). In addition, the bases in probes may be joined by a linkage other than a phosphodiester bond, so long as the bond does not interfere with hybridization. Thus, probes may be peptide nucleic acids in which the constituent bases are joined by peptide bonds rather than phosphodiester linkages.

"Target nucleic acid" refers to a nucleic acid (often derived from a biological sample), to which the probe is designed to specifically hybridize. It is either the presence or absence of the target nucleic acid that is to be detected, or the amount of the target nucleic acid that is to be quantified. The target nucleic acid has a sequence that is complementary to the nucleic acid sequence of the corresponding probe directed to the target. The term target nucleic acid may refer to the specific subsequence of a larger nucleic acid to which the probe is directed or to the overall sequence (e.g., gene or mRNA) whose expression level it is desired to detect. The difference in usage will be apparent from context.

An "array" may comprise a solid support with peptide or nucleic acid probes attached to said support. Arrays typically comprise a plurality of different nucleic acids or peptide probes that are coupled to a surface of a substrate in different, known locations. These arrays, also described as "microarrays" or colloquially "chips" have been generally described in the art, for example, U.S. Pat. Nos. 5,143,854, 5,445,934, 5,744,305, 5,677,195, 6,040,193, 5,424,186 and Fodor et al., Science, 251:767-777 (1991). Each of which is incorporated by reference in its entirety for all purposes. These arrays may generally be produced using mechanical synthesis

methods or light directed synthesis methods which incorporate a combination of photolithographic methods and solid phase synthesis methods. Techniques for the synthesis of these arrays using mechanical synthesis methods, such as ink jet, channel block, flow channel, and spotting methods which are described in, e.g., U.S. Pat. Nos. 5,384,261, and 6,040,193, which are incorporated herein by reference in their entirety for all purposes. Although a planar array surface is preferred, the array may be fabricated on a surface of virtually any shape or even a multiplicity of surfaces. Arrays may be peptides or nucleic acids on beads, gels, polymeric surfaces, fibers such as fiber optics, glass or any other appropriate substrate. *See* U.S. Patent Nos. 5,744,305, 5,770,358, 5,789,162, 5,708,153, 6,040,193 and 5,800,992, which are hereby incorporated in their entirety for all purposes. Arrays may be packaged in such a manner as to allow for diagnostics or other manipulation of in an all inclusive device. *See* for example, US Patent Nos. 5,856,174 and 5,922,591, and 5,945,334, which are incorporated herein in their entirety by reference for all purposes. *See* also U.S. patent application Serial No. 09/545,207 which is incorporated herein in its entirety for all purposes for additional information concerning arrays, their manufacture, and their characteristics. It is hereby incorporated by reference in its entirety for all purposes.

II. Probe Selection Systems

As will be appreciated by one of skill in the art, the present invention may be embodied as a method, data processing system or program products. Accordingly, the present invention may take the form of data analysis systems, methods, analysis software, etc. Software written according to the present invention is to be stored in some form of computer readable medium, such as memory, or CD-ROM, or transmitted over a network, and executed by a processor. For a

description of basic computer systems and computer networks, *See*, e.g., Introduction to Computing Systems: From Bits and Gates to C and Beyond by Yale N. Patt, Sanjay J. Patel, 1st edition (January 15, 2000) McGraw Hill Text; ISBN: 0072376902; and Introduction to Client/Server Systems : A Practical Guide for Systems Professionals by Paul E. Renaud, 2nd edition (June 1996), John Wiley & Sons; ISBN: 0471133337.

Computer software products may be written in any of various suitable programming languages, such as C, C++, Fortran and Java (Sun Microsystems). The computer software product may be an independent application with data input and data display modules. Alternatively, the computer software products may be classes that may be instantiated as distributed objects. The computer software products may also be component software such as Java Beans (Sun Microsystems), Enterprise Java Beans (EJB), Microsoft® COM/DCOM, etc.

FIG. 1 illustrates an example of a computer system that may be used to execute the software of an embodiment of the invention. FIG. 1 shows a computer system 1 that includes a display 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have one or more buttons for interacting with a graphic user interface. Cabinet 7 houses a CD-ROM or DVD-ROM drive 13, system memory and a hard drive. *See* FIG. 2 which may be utilized to store and retrieve software programs incorporating computer code that implements the invention, data for use with the invention and the like. Although a CD 15 is shown as an exemplary computer readable medium, other computer readable storage media including floppy disk, tape, flash memory, system memory, and hard drive may be utilized. Additionally, a data signal embodied in a carrier wave (*e.g.,* in a network including the Internet) may be the computer readable storage medium.

FIG. 2 shows a system block diagram of computer system 1 used to execute the software of an embodiment of the invention. As in FIG. 1, computer system 1 includes monitor 3, keyboard 9, and mouse 11. Computer system 1 further includes subsystems such as a central processor 51, system memory 53, fixed storage 55 (*e.g.*, hard drive), removable storage 57 (*e.g.*, CD-ROM), display adapter 59, sound card 61, speakers 63, and network interface 65. Other computer systems suitable for use with the invention may include additional or fewer subsystems. For example, another computer system may include more than one processor 51 or a cache memory. Computer systems suitable for use with the invention may also be embedded in a measurement instrument.

III. Methods for Predicting Quality Scores of Probes

In a preferred embodiment, arrays of oligonucleotides or peptides, for example, are formed on the surface by sequentially removing a photoremovable group from a surface, coupling a monomer to the exposed region of the surface, and repeating the process. These techniques have been used to form extremely dense arrays of oligonucleotides, peptides, and other materials. The synthesis technology associated with this invention has come to be known as "VLSIPS™ " or "Very Large Scale Immobilized Polymer Synthesis" technology and is further described below.

Additional techniques for forming and using such arrays are described in U.S. Patent Nos. 5,384,261, and 6,040,193 which are also incorporated by reference for all purposes. Such techniques include systems for mechanically protecting portions of a substrate (or chip), and selectively deprotecting/coupling materials to the substrate. Still further techniques for array

synthesis are provided in U.S. Application No. 08/327,512, also incorporated herein by reference for all purposes.

Nucleic acid probe arrays have found wide applications in gene expression monitoring, genotyping and mutation detection. For example, massive parallel gene expression monitoring methods using nucleic acid array technology have been developed to monitor the expression of a large number of genes (e.g., U.S. Patent Numbers 5,871,928, 5,800,992 and 6,040,138; de Saizieu *et al.*, 1998, Bacteria Transcript Imaging by Hybridization of total RNA to Oligonucleotide Arrays, NATURE BIOTECHNOLOGY, 16:45-48; Wodicka *et al.*, 1997, Genome-wide Expression Monitoring in Saccharomyces cerevisiae, NATURE BIOTECHNOLOGY 15:1359-1367; Lockhart *et al.*, 1996, Expression Monitoring by Hybridization to High Density Oligonucleotide Arrays. NATURE BIOTECHNOLOGY 14:1675-1680; Lander, 1999, Array of Hope, NATURE-GENETICS, 21 (suppl.), at 3, all incorporated herein by reference for all purposes). Hybridization-based methodologies for high throughput mutational analysis using high-density oligonucleotide arrays (DNA chips) have been developed, *See* Hacia *et al.*, 1996, Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-color fluorescence analysis. Nat. Genet. 14:441-447, Hacia *et al.*, New approaches to BRCA1 mutation detection, Breast Disease 10:45-59 and Ramsey 1998, DNA chips: State-of-Art, Nat Biotechnol. 16:40-44, all incorporated herein by reference for all purposes). Oligonucleotide arrays have been used to screen for sequence variations in, for example, the *CFTR* gene (U.S. Patent Number 6,027,880, Cronin et al., 1996, Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. Hum. Mut. 7:244-255, both incorporated by reference in their entireties), the human immunodeficiency virus (HIV-1) reverse transcriptase and protease genes (U.S. Patent Number 5,862,242 and Kozal et al., 1996, Extensive polymorphisms observed in

HIV-1 clade B protease gene using high density oligonucleotide arrays. Nature Med. 1:735-759, both incorporated herein by reference for all purposes), the mitochondrial genome (Chee et al., 1996, Accessing genetic information with high density DNA arrays. Science 274:610-614) and the BRCA1 gene (U.S. Patent Number 6,013,449, incorporated herein by reference for all purposes).

In one aspect of the invention, a physical model that is based on the thermodynamic properties of the sequence is used to predict the array-based hybridization intensities of the sequence. Hybridization propensities may be described by energetic parameters derived from the probe sequence, and variations in hybridization and chip manufacturing conditions will result in changes in these parameters that can be detected and corrected. U.S. Patent Application Serial Number 09/718,29, docket Number 3367, filed on November 21, 2000, and incorporated herein by reference, discloses methods for predicting nucleic acid hybridization affinity.

The values of weight coefficients in the physical model may be determined by empirical data because these values are influenced by assay conditions, which include hybridization and target fragmentation, and probe synthesis conditions, which include choice of substrates, coupling efficiency, etc.

In one embodiment (FIG. 3), a model experimental system is used to generate empirical data and a computational model is used to process these data to solve for the weight coefficients of the physical model. These solved weight coefficients are in turn placed back into the physical model, enabling it to predict the hybridization behaviors of new sequences.

The interaction between a probe and its target is described in FIG. 4. Basically, a target (T) hybridizes to its complementary probe (P) to form a probe-target duplex (P•T) (FIG. 4), and the reaction is accompanied with favorable free energy change (FIG. 4). The amplitude

of the free energy change (ΔG) determines the stability of probe-target duplex. The duplex stability can be described by equilibrium constant (K_s), which is sequence-dependent. The relationship between K_s and ΔG may be given by Boltzmann's equation:

$$K_s = \frac{k_{on}}{k_{off}} = e^{-\Delta G / RT} \qquad \text{[Equation 1]}$$

where k_{on} and k_{off} are the rate constants for association and dissociation, respectively, of the probe-target duplex, R is the gas constant and T is the absolute temperature. According to Equation 1, ΔG is a function of the sequence. The dependence of ΔG on probe sequence can be quite complicated, but relatively simple models for ΔG have yielded good results.

There are a number of ways to establish the relationship between the sequence and ΔG. In preferred embodiments, one model (equation 2), shown in U.S. Patent Application Serial Number 09/718,29, previously incorporated by reference is shown below:

$$\Delta G_{seq} = \sum_{i=1}^{3N} P_i S_i \qquad \text{[Equation 2]}$$

or

$$\Delta G_{seq} = \sum_{i=1}^{3N} P_i S_i + C \qquad \text{[Equation 3]}$$

where N is the length (number of bases) of a probe. P_i is the value of the ith parameter which reflects the ΔG of a base in a given sequence position relative to a reference base in the same position. In preferred embodiments, the reference base is A. In this case, the Pi's will be the free energy of a base in a given position relative to base A in the same position. FIG. 5 shows an example of how the value of S_i is determined based upon the sequence of a probe. In this example, a probe, GTCA has N=4 and thus, it has 3x4=12 S_i values. Each probe base position

has three S values, each for a different possible base. In this example, possible bases are evaluated in the sequence of C, G, and T (A is the reference base). However, one of skill in the art would appreciate that the assignment of this particular of base sequence is arbitrary. Alternatively evaluation sequence, such as G, C, and T may also be used as long as the same scheme is used for model building and for hybridization affinity prediction.

Based on the simple hybridization scheme described in FIG. 4, the hybridization intensity is proportional to the concentration of probe-target duplex, where C_0 is constant (Equation 4). Under equilibrium condition, the intensity is directly related to ΔG (Equation 5). This relationship is also expressed in natural logarithm form, where C_1 and C_2 are constants (Equation 7) and Equation 6 also holds for approaching equilibrium cases. According to Equation 2, the relationship between intensity and probe sequence is described in Equation 7 and 8:

$$I = C_0\,[P{\bullet}T] \qquad \text{[Equation 4]}$$

$$[P{\bullet}T] = K_s[P][T] = e^{-\Delta G/RT}\,[P][T] \qquad \text{[Equation 5]}$$

$$\mathrm{LnI} = -\Delta G/RT + \mathrm{Ln}\{C_0[P][T]\} \qquad \text{[Equation 6]}$$

C_2

$$LnI = C_1\sum_{i=1}^{3N} P_iS_i + C_2\text{, where } C_2 =$$

$$\mathrm{Ln}\{C_0[P][T]\} \text{ and } C_1 = -1/RT \qquad \text{[Equation 7]}$$

or

$$LnI = \sum_{i=1}^{3N} C_1P_iS_i + C_2 = \sum_{i=1}^{3N} W_iS_i + C_2 \qquad \text{[Equation 8]}$$

where $W_i=C_1P_i$. The following is a linear regression model for probes of N bases in length using a training data set that contains intensity values of M probes.

$$Ln(I_1) = W_1 S_{11} + W_2 S_{21} + \ldots W_{3N} S_{3N1}$$
$$Ln(I_2) = W_1 S_{12} + W_2 S_{22} + \ldots W_{3N} S_{3N2}$$
$$\vdots$$
$$Ln(I_1) = W_1 S_{11} + W_2 S_{12} + \ldots W_{3N} S_{3N1}$$

Hybridization intensities (relative to a reference base, such as an A) for each type of bases can be solved at each position in the probe sequence may be predicted. Multiple linear regression analysis is well known in the art. *See*, for example, the electronic statistic book (http://www.statsoftinc.com/textbook/stathome.html); Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill, both incorporated by reference for all purposes. Computer software packages, such as SAS, SPSS, and MatLib 5.3 provide multiple linear regression functions. In addition, computer software code examples suitable for performing multiple linear regression analysis are provided in, for example, the Numerical Recipes (NR) books developed by Numerical Recipes Software and published by Cambridge University Press (CUP, with U.K. and U.S. web sites).

In a preferred embodiment, a set of probes of different sequences (probes 1 to M) is used as probes in experiments(s). Hybridization affinities (relative ΔG or Ln (I)) of the probes with their target are experimentally measured to obtain a training data set. (*See* example section *infra.*) Multiple linear regression may be performed using hybridization affinities as I $[I_1 \ldots I_m]$ to obtain a set of weight coefficients: $[W_1 \ldots W_N]$. The weight coefficients are then used to predict the hybridization affinities using Equation 7. FIG. 6A shows relative predicted ΔG at every base position in a probe of 25 bases in an exemplary experiments. (*See* the example section *infra* for a detailed description of experimental conditions.)

In addition, in some embodiments, by using intensities derived from mismatch probes that are probes designed to contain one or more mismatch bases from a reference probe, a set of weight coefficients may be obtained to predict the mismatch intensity using perfect match probe sequence. FIG. 6B shows an example for predicting mismatch hybridization affinity at center base position.

Since other interactions such as probe self-folding, probe-to-probe interaction, target self-folding and target-to-target interaction also interfere with the probe-target duplex formation, their contributions to the values of the weight coefficients may also be considered. FIG. 7 shows an overall equilibrium scheme including the formation of a probe-target duplex (PT), probe self-folding (P_F) and probe dimerization (PP). Probe folding renders the probe unavailable for binding with the target. Probe dimerization renders two probes unavailable for binding with the target. In some embodiments, the hybridization affinity prediction model accounts for probe folding and probe dimerization:

$$\Delta G^0_{overall} = -W_d \Delta G^0_d + W_{PF} \Delta G^0_{PF} + W_{PP} \Delta G^0_{pp} \qquad \text{[Equation 9]}$$

$$\ln I = C_1 \Delta G^0_{overall} + C_2 \qquad \text{[Equation 10]}$$

where W_d is the weight for sequence based probe affinity; W_{PF} is the weight for probe formation and W_{PP} is the weight for probe dimerization. Any methods that are capable of predicting probe folding and/or probe dimerization are suitable for at least some embodiments of the invention for predicting the hybridization intensity in at least some embodiments of the invention. In a particularly preferred embodiment, Oligowalk (available at http://rna.chem.rochester.edu/RNAstructure.html, last visited Nov. 3, 2000) may be used to predict probe folding.

One important criterion of probe selection for a quantitative gene expression assay is that hybridization intensities of the selected probes must correspond to target concentration changes. In some embodiments, the relationship between concentrations and intensities of a probe is modeled as:

$$Ln(I) = SLnC + LnK_{app} \qquad \text{[Equation 11]}$$

or

$$I = K_{app}C^S \qquad \text{[Equation 12]}$$

where I is intensity; K_{app} is apparent affinity constant; C is concentration of the target; and S is an empirical value corresponding to the slope of the line relates LnI and LnC ($0<S<1$). (*See* FIG. 8.)

Equation 12 describes the relationship between hybridization intensities of probes and target concentration. For example, when S is equal to 1, the intensities of a probe linearly correspond to its target concentration (FIG. 8). Thus, based on the S values of the probe, one can select probes that have good concentration dependence. FIG. 9A shows the polynomial relationship between S and LnK_{app}, indicating that when the value of LnK_{app} increases to a certain level the value of S reaches a plateau before starting to decrease. This relationship allows the identification of not only low hybridization affinity probes (FIG. 9B, bottom lines) but also GC-rich probes that have high affinity but bind to both specific and non-specific targets (FIG. 9B, top line). These GC-rich probes have high intensities, but the intensities maintain constant when target concentration changes (FIG. 9B, top line). Therefore, these probes have small slopes. In some embodiments, linear regression modeling alone will not identify probes with a high propensity to saturate. That is because the linear model for each target concentration will predict the intensity that a probe would have had if it could bind to unlimited amount of target.

Therefore, the predicted slope can be quite high when the observed slope is low (FIG. 24, top). The well-behaved relationship between predicted LnK_{app} and observed slope allows filtering probes with a high propensity to saturate based on the predicted LnK_{app} for the given probe. If LnK_{app} is above a cutoff value (e.g., 5, 6, 7, 8, 9, or 10, FIG. 9), then the probe is effectively filtered as a candidate for probe selection. FIG. 24 (middle) shows the predicted slope profiles after filtering as well as the significant improvement in the overall correlation after these regions are removed.

IV. Methods and Software for Selecting Probes

FIG. 10 shows a computer-implemented process for selecting probe sequences from a pool of candidate probes. In this particularly embodiment, the sequences of a pool of candidate oligonucleotide probe are processed by a quality predictor (101). Throughout this application, the term probe may refer to the sequence of a probe. The pool of candidate oligonucleotide probes may be all possible probes against a particular target or targets. Typically, oligonucleotide probes are at least 10, 15, 20, 25 and 30 bases in length. Polynucleotide probes can be more than 10, 20, 25, 30, 100, 200, 500, 1000, or 5000 bases in length. FIG. 11 illustrates a complete pool of candidate oligonucleotide probes (unfiled rectangle boxes) against a target (black rectangle box). Each of the probes is designed to be complementary to the target sequence. In this particular embodiment, the oligonucleotides are 25mers. The first probe is complementary to bases 1-25 (from the 5' end) of the target sequence. The second probe is complementary to bases 2-26 and so on. While a complete pool is often desirable, it is not necessary to have a complete pool for at least some embodiments of the invention. In some cases, filters may be used to remove some of the probes from the pool.

The input to the quality predictor (FIG. 10, 101) is the sequence of a pool of candidate probes. One of skill in the art would appreciate that the format of input is not critical. In some embodiments, the probe sequences may be inputted from one or a number of probe sequence files. The file(s) may be plain text file(s), in the FASTA format or other suitable file format. Alternatively, the input may be a stream from other sources such as a data pocket stream from a remote networked computer.

The quality predicator is a software module that calculates quality scores (the term score refers to any qualitative and quantitative values with regard to desired properties of a probe) for probes based upon the sequences of probes. In some embodiments, the quality score may include predicted values such as perfect match intensity, mismatch intensity and/or slope.

Probe selection module (103) selects probes based upon their scores. In preferred embodiments, the quality scores are combined to obtain a unified score. In some cases, the unified quality score is the simple summation of quality scores (e.g., Unified Quality Score=Perfect Match Intensity+Mismatch Intensity+Slope). The selection of probes may be based upon the scores only. For example, if certain number of probes are desired, the probes with the highest scores are selected until enough number of probes are selected. Alternatively, a threshold-unified score may be established. Probes that have scores higher than the threshold score are selected.

In preferred embodiment, the goal of probe selection step is to find the best probes to represent a sequence. The probe selection software module takes a set of probes and a set of quality measures for each probe. It then implements an optimization algorithm to find the best n probes, spread out across the gene. Methods for probe selection using optimization algorithm is

described in U.S. Provisional Application Number 60/252,617, incorporated herein by reference in its entirety for all purposes.

FIG. 12 shows another embodiment of the computer implemented probe selection process of the invention, target sequences are inputted to a candidate probe generator (121) which produce either all possible probes of certain length or a subset of the all possible probes. The candidate probe sequences are fed to the quality score predictor (122) for calculating quality measures (scores, e.g., perfect match intensity, mismatch intensity and/or slope). The candidate probe sequences are also fed to a 3' bias score predictor (123) to obtain 3' bias scores that indicates the distance of probe sequence from the 3' end of target sequence. Since the current target preparation method is 3' biased, it is important to select probes that fall into range where its target will be made. The probe sequences may optionally be inputted into a cross hybridization score predictor (124) to calculate cross hybridization scores. The quality scores, 3' bias scores and/or cross hybridization score are combined by a probe score calculator module (125) to produce a unified score. A probe selection module (126) picks the probes with the score.

FIG. 13 shows a complete computer implemented probe selection process. In this preferred embodiment, target sequences (131) are used to generated a pool of candidate probes. The probe sequences are stored in a FASTA sequence file. A sequence file splitter (132) divides probe sequences to .seq file which store one sequence per file. The .seq files are processed using a OligoWalker batch tool (133) to produce a .rep file, one for each probe sequence. The .rep files contain ΔG values for the probes. The rep files are inputted into a quality predictor (134). The quality predictor is based upon a multiple linear regression models derived from experiment data using, for example, yeast test chips. (*See* also, example section below) (1310). The quality

predictor calculates quality scores (measures, perfect match intensity, mismatch intensity and slope) as described above in section II. The rep file is also inputted into a 3' bias score predictor (135) to estimate 3' bias scores for the probes.

The multiple probe FASTA sequence file is also inputted into a cross hybridization predictor (136) to predict a cross hybridization score. The cross hybridization score predictor is based upon models (such as multiple linear regression models) derived from experiment data (1311). In some embodiments, cross hybridization may also be evaluated by pruning probe sequences against a human genome data base (1312) which may be residing locally, in a local area network or in a remote site such as the Genbank (http://www.ncbi.nlm.nih.gov).

The quality measures, 3' bias scores and cross hybridization scores are combined by the probe score calculator (137) to produce a unified score for each probe. The combined score is then used for selecting probes (138). The probe selection module takes a set of probes and a set of quality measures for each probe. It then implements a dynamic programming algorithm to find the best n probes, spread out across the gene. The selected probe sequences are stored in .101 files (139).

The following tables describe the various software modules in the examplary embodiments described in FIG. 13.

1 Multiple linear regression modeling tool

Description	Calculates the weights for the regression model. Its is a one time calculation. The results of the calculations will be used every time a new chip is designed.
Input	Yeast Test Chip, available from Affymetrix, Santa Clara, CA

Output	Multiple linear regression models, a set of weights.
Part of chip design	In this embodiment, it is not part of the software package for chip design. It is used as one time external process. However, in other exemplary embodiments, it may also become part of the software.

2. Sequence file splitter

Description	Splits a FASTA file of sequences into several sequence files one for each sequence in the instruction file. If max files in folder is greater than 0, subfolders are created in the output path. Each subfolder get up to the maximum files specified.
Input	• FASTA file • Instructions file • Output path • Max files in one folder
Language / Tool	Java

3. Oligo Walk batch tool

Description	Runs Oligo Walk in batch mode. Oligo Walk produces a .rep file for each sequence. The .rep file contains a delta G values for each probe
Input	Batch of .seq files
Output	.rep file. The .rep file identifies a probe by a number and a sequence. The sequence is a reverse complement of the 25-mer it represents on the input sequence. The number is the beginning of the probe.

Part of chip design	Yes
Language / Tool	Microsoft ® Visual Basic

4. Quality predictor

Description	Takes in the MLR model measures and delta G values from Oligo Walk and produces 3 quality measures, perfect match intensity, mismatch intensity and slope.
Input	.rep file produced from Oligo Walk
Output	3 Quality measures for each probe. The probe is described as in the input format.
Part of chip design	Yes
Language / Tool	C

5. Cross Hyb Modeling Tool

Description	Analyzes the results of the yeast cross hyb chip to create a model for predicting the cross hyb score for a probe, based on the number of mismatches and positions of mismatches with 1 or more matching sequences.
Input	Results from the cross hyb chip
Output	A model that relates number of mismatches and positions of mismatches to a cross hyb score.
Part of chip	In some embodiments, it is not part of the chip design package.

design	Alternatively, it can be part of the package.

6. Cross Hyb Score Predictor

Description	Predicts a cross hyb score for a given set of probes. Its does so by matching the given probes with a genome and assigns a numeric score using the cross hyb models.
Input	▪ Cross hyb models ▪ A genome ▪ Set of probes
Output	List of probes and corresponding cross hyb scores
Part of chip design	Yes

7. 3' Bias Score Predictor

Description	Predicts the 3' bias score for a given set of probes. Earlier it was believed that most sequences have a sigmoid graph for the 3' bias. But, recently used sequences do not always follow the pattern Therefore, it is important to first study the 3' bias effect and then design a measurement model.
Input	▪ Set of probes
Output	List of probes and corresponding 3' bias scores
Part of chip design	Yes

8. Probe Score Calculator

Description	Given a set files with probe information and scores, this program matches each probe in each sequence and calculates 1 unified score for each probe.

Input	▪ Set of probes
	▪ Set of measures for each probe, each in a different file(s)
	    ▪ 3 quality scores (probes defined in OligoWalk format)
	    ▪ cross hyb score (probes defined in chip design format)
	    ▪ 3' bias score (probes defined in chip design format)

Output	List of probes with a corresponding score.
Part of chip design	Yes

9. Probe selection algorithm

Description	Finds the best probes to represent a sequence. It takes a set of probes and a set of quality measures for each probe. It then implements a dynamic programming algorithm to find the best n probes, spread out across the gene.

Input	▪ Set of probes
	▪ Set of measures for each probe
	  ▪ 3 quality scores
	  ▪ cross hyb score
	  ▪ 3' bias score
	▪ Number of probes to choose

Output	.1lq file
Part of chip design	Yes
Language / Tool	C

10. Algorithm Test Tool

Description	Tests the new probe selection algorithm. The probe selection algorithm is used to select probes for the known, Yeast test chip. The selected probes are analyzed for their intensity, slope and discrimination values on the yeast test chip.
Input	▪ Probes selected for the sequences on the yeast test chip ▪ Results from the yeast test chip

V. Examples

The following examples demonstrate the effectiveness of the methods of the invention for predicting hybridization intensities and for selecting oligonucleotide probes for gene expression monitoring.

A. Example 1: Prediction of Hybridization Intensities of Probes Against Yeast Genes

FIG. 14 shows the overall process of the experiments. Yeast was used as a model system for this experiment because the yeast genome had been sequenced. Arrays containing nucleic acid probes complementary to yeast genes are commercially available from Affymetrix (Santa Clara, California). Genes were selected to cover sequence complexity such as GC content, secondary structure, Motif and gene clusters. Twenty probe pairs (perfect match and mismatch probes) were selected to cover entire sequence of one of the 112 selected yeast genes. The probes are synthesized in situ on glass substrate using photo-directed synthesis method that was disclosed in, for example, U.S. Patent Nos. 5,384,261, and 5,744,305, 5,445,934 and 6,040,138.

One hundred and twelve yeast clones representing the 112 genes were randomly divided into 14 groups (FIG. 15). Labeled targets prepared from these clones were used as spikes for 14 experiments at various concentration levels from 0pM to 1024pM. In some experiments, the spikes derived from yeast gene clones were combined with labeled nucleic acid representing human complex background. A 14 x 14 Latin square design (FIG. 16) was employed. The numbers in the table indicates the concentration used (pM). For each experiment, 14 groups of genes at 14 different concentrations were pooled together and hybridized to an oligonucleotide probe array. For each Latin Square 14 oligonucleotide probe array hybridization experiments were performed. FIG. 17 shows experiments conducted.

Cross-validation (FIG. 18) was used to evaluate the prediction. The cross-validation process held one gene for test and used the other 111 genes to solve the weight coefficients that in turn were used to predict intensities for the test genes, as described in FIG. 14. The correlation between the predicted and measured intensity for one test gene (YDR113C) is shown in FIG. 19A and FIG. 19B shows the correlation against target sequence, where lines represent the predicted values and dots represent the observed values. The correlation of the predicted and measured values for perfect match (PM) and mismatch (MM) probes is also demonstrated in FIG. 20A and 20B respectively, where lines represent the predicted values and dots represent the observed values for gene YGR109C.

FIG. 21 shows predicted intensity versus actual intensity at various target spike concentrations, where lines indicate the predicted values and dots represent observed values. FIG. 22 shows correlation coefficients between predicted and observed intensity (LnI) as function of concentration, where top and bottom lines represent perfect matches and mismatches, respectively. The high correlation (0.85) holds for 4000-fold concentration range (FIG. 22), and the results demonstrate that the methods of invention are able to predict probe behaviors through a wide dynamic range.

FIG. 23 shows predicted versus observed intensities when the target transcripts were derived from genes in the wrong orientation, which results no complimentary target was generated for the probes. As shown in FIG. 24, predicted intensities (lines) had no correlation with observed intensity (dots) because right target is absent. The result indicates the prediction method is accurate and specific.

FIG. 24 shows predicted slope versus observed slope. In some regions in the FIG. 24(top), the values of predicted slope (lines) can be quite high when the values of observed slope

(dots) because of the saturated probes in those regions. According to Equation 12 and FIG. 9, the saturated probes can be identified and removed. FIG. 24 (middle) shows the predicted slope profiles after filtering the saturated probes and the significant improvement in the overall correlation after these regions are removed.

B. Example 2: Prediction of Hybridization Intensities of Probes from Human Genes

This example demonstrates that weight coefficients obtained from the model yeast experiment system is also able to predict the intensities on the human gene expression chip and the predicted intensities (left bar) are highly correlated with observed intensities (right bar) at each probe position as indicated by x-axis. The correlation is shown in FIGs. 25 A-E. Typically, the correlation coefficients ranged from 0.45-0.83. The distribution of the correlation coefficients are shown in FIG. 26. These results demonstrate that the probe selection model may be generalized to different organisms such as mammals, plants,

C. Example 3: Probe Selection

This example demonstrates that the model-based probe selection method and software may provide improvement over current probe selection methods. FIG. 27 shows intensity values of sixteen probes (open squares) selected for the Yer161c gene based upon quality scores and using dynamic programming. FIG. 27 also shows that the sixteen selected probes (open squares) are spaced along sequence. FIG. 28 shows a comparison of average intensity difference (between perfect match and mismatch) values of probe selected by various methods for all yeast test genes. Probes selected randomly (diamonds) were similar to those selected according empirical rules (squares). The model based selection method (triangles) improved average

intensity difference values. The result indicates the model-selected probes have high sensitivity and specificity.

CONCLUSION

The present invention provides methods and computer software products for predicting nucleic acid hybridization affinity, detecting mutation, selecting better-behaved probes, and improving probe array manufacturing quality control. It is to be understood that the above description is intended to be illustrative and not restrictive. Many variations of the invention will be apparent to those of skill in the art upon reviewing the above description. By way of example, the invention has been described primarily with reference to the use of a high density oligonucleotide array, but it will be readily recognized by those of skill in the art that the methods may be used to predict the hybridization affinity of other immobilized probes, such as probes that are immobilized in or on optical fibers or other supports by any deposition methods. The basic methods and computer software of the invention may also be used to predict solution-based hybridization. The scope of the invention should, therefore, be determined not with reference to the above description, but should instead be determined with reference to the appended claims, along with the full scope of equivalents to which such claims are entitled.

All references cited herein are incorporated herewith by reference for all purposes.

CLAIMS

We claim:

1. A computer implemented method for selecting oligonucleotide probes comprising:

   a) predicting hybridization intensities of a plurality of candidate probes;

   b) predicting quantitative responses of said candidate probes to the amount of their targets; and

   c) selecting said probes from said candidate probes according to said hybridization intensities and said quantitative response.

2. The method of Claim 1 wherein said quantitative response is the slope of the response curve of said probe.

3. The method of Claim 2 wherein said hybridization intensity (I) is determined using the equation:

$$Ln(I) = \sum_{i=1}^{3N} W_i S_i + C_2$$

Or

$$Ln(I) = \sum_{i=1}^{3N} W_i S_i$$

wherein said W_i is a weight coefficient; S_i is a functional of said sequence of said probe; N is the number of bases of said probe; and C_2 is a constant.

4. The method of Claim 3 wherein said weight coefficient is determined using multiple linear regression analysis.

5. The method of Claim 4 further comprising predicting mismatch hybridization intensities of corresponding mismatch probes of said candidate probes and wherein said selecting step is also based upon said mismatch hybridization intensities.

6. The method of Claim 5 wherein said mismatch probes are different from their corresponding candidate probes in one base pair in the middle of their sequences.

7. The method of Claim 6 wherein said mismatch hybridization intensities are predicted according to the sequences of said candidate genes.

8. The method of Claim 3 further comprising filtering out a subset of said candidate probes, wherein said subset probes have apparent affinity constant above a threshold.

9. The method of Claim 8 wherein the threshold is above 5 for ln (apparent affinity constant).

10. The method of Claim 9 wherein the threshold is above 6.

11. The method of Claim 10 wherein the threshold is above 7.

12. The method of Claim 7 wherein mismatch hybridization intensities are determined according to the following equation:

$$Ln(I) = \sum_{i=1}^{3N} W_i' \; S_i + C_2' \quad \text{or}$$

$$Ln(I) = \sum_{i=1}^{3N} W_i' \; S_i$$

wherein said W'_i is a weight coefficient; S_i is a functional of said sequence of said probe; N is the number of bases of said probe; and C_2' is a constant.

13. The method of Claim 12 wherein said selecting step comprises calculating a unified quality score based upon predicted hybridization intensities.

14. A computer software product for selecting oligonucleotide probes comprising:

computer program code for predicting hybridization intensities of a plurality of candidate probes;

computer program code for predicting quantitative responses of said candidate probes to the amount of their targets;

computer program code for selecting said probes from said candidate probes according to said hybridization intensities and said quantitative response; and

a computer readable media for storing said computer program codes.

15. The computer software product of Claim 14 wherein said quantitative response is the slope of the response curve of said probe.

16. The computer software product of Claim 15 wherein said hybridization intensity (I) is determined using the equation:

$$Ln(I) = \sum_{i=1}^{3N} W_i S_i + C_2$$

Or

$$Ln(I) = \sum_{i=1}^{3N} W_i S_i$$

wherein said W_i is a weight coefficient; S_i is a functional of said sequence of said probe; N is the number of bases of said probe; and C_2 is a constant.

17. The computer software product of Claim 16 wherein said weight coefficient is determined using multiple linear regression analysis.

18. The computer software product of Claim 17 further comprising computer program code for predicting mismatch hybridization intensities of corresponding mismatch probes of said candidate probes and wherein said selecting step is also based upon said mismatch hybridization intensities.

19. The computer software product of Claim 18 wherein said mismatch probes are different from their corresponding candidate probes in one base pair in the middle of their sequences.

20. The computer software product of Claim 19 wherein said mismatch hybridization intensities are predicted according to the sequences of said candidate genes.

21. The computer software product of Claim 20 wherein mismatch hybridization intensities are determined according to the following equation:

$$Ln(I) = \sum_{i=1}^{3N} W_i^{'} \ S_i + C_2^{'} \quad \text{or}$$

$$Ln(I) = \sum_{i=1}^{3N} W_i^{'} \ S_i$$

wherein said W'_i is a weight coefficient; S_i is a functional of said sequence of said probe; N is the number of bases of said probe; and $C_2^{'}$ is a constant.

22. The computer software product of Claim 14 further comprising computer program code of filtering out a subset of said candidate probes, wherein said subset probes have apparent affinity constant above a threshold.

23. The computer software product of Claim 22 wherein the threshold is above 5 for ln (apparent affinity constant).

24. The computer software product of Claim 23 wherein the threshold is above 6.

25. The computer software product of Claim 24 wherein the threshold is above 7.

26. The computer software of Claim 21 wherein said computer program code for selecting comprises computer program code for calculating a unified quality score based upon predicted hybridization intensities.

27. A system for selecting nucleic acid probes, comprising:

a processor; and

a memory being coupled to the processor, the memory storing a plurality machine instructions that cause the processor to perform a plurality of logical steps when implemented by the processor, said logical steps including:

predicting hybridization intensities of a plurality of candidate probes;

predicting quantitative responses of said candidate probes to the amount of their targets; and

selecting said probes from said candidate probes according to said hybridization intensities and said quantitative response.

28. The system of Claim 27 wherein said quantitative response is the slope of the response curve of said probe.

29. The system of Claim 28 wherein said hybridization intensity (I) is determined using the equation:

$$Ln(I) = \sum_{i=1}^{3N} W_i S_i + C_2$$

OR

$$Ln(I) = \sum_{i=1}^{3N} W_i S_i$$

wherein said W_i is a weight coefficient; S_i is a functional of said sequence of said probe; N is the number of bases of said probe; and C_2 is a constant.

30. The system of Claim 29 wherein said weight coefficient is determined using multiple linear regression analysis.

31. The system of Claim 27 wherein said logic steps further comprises predicting mismatch hybridization intensities of corresponding mismatch probes of said candidate probes and wherein said selecting step is also based upon said mismatch hybridization intensities.

32. The system of Claim 31 wherein said mismatch probes are different from their corresponding candidate probes in one base pair in the middle of their sequences.

33. The system of Claim 32 wherein said mismatch hybridization intensities are predicted according to the sequences of said candidate genes.

34. The system of Claim 33 wherein mismatch hybridization intensities are determined according to the following equation:

$$Ln(I) = \sum_{i=1}^{3N} W_i^{'} S_i + C_2^{'} \quad \text{or}$$

$$Ln(I) = \sum_{i=1}^{3N} W_i^{'} S_i$$

wherein said W'_i is a weight coefficient; S_i is a functional of said sequence of said probe; N is the number of bases of said probe; and $C_2^{'}$ is a constant.

35. The system of Claim 27 wherein said logic steps further compries filtering out a subset of said candidate probes, wherein said subset probes have apparent affinity constant above a threshold.

36. The system of Claim 35 wherein the threshold is above 5 for ln (apparent affinity constant).

37. The system of Claim 35 wherein the threshold is above 6.

38. The system of Claim 35 wherein the threshold is above 7.

39. The system of Claim 34 wherein said selecting step comprises calculating a unified quality score based upon predicted hybridization intensities.

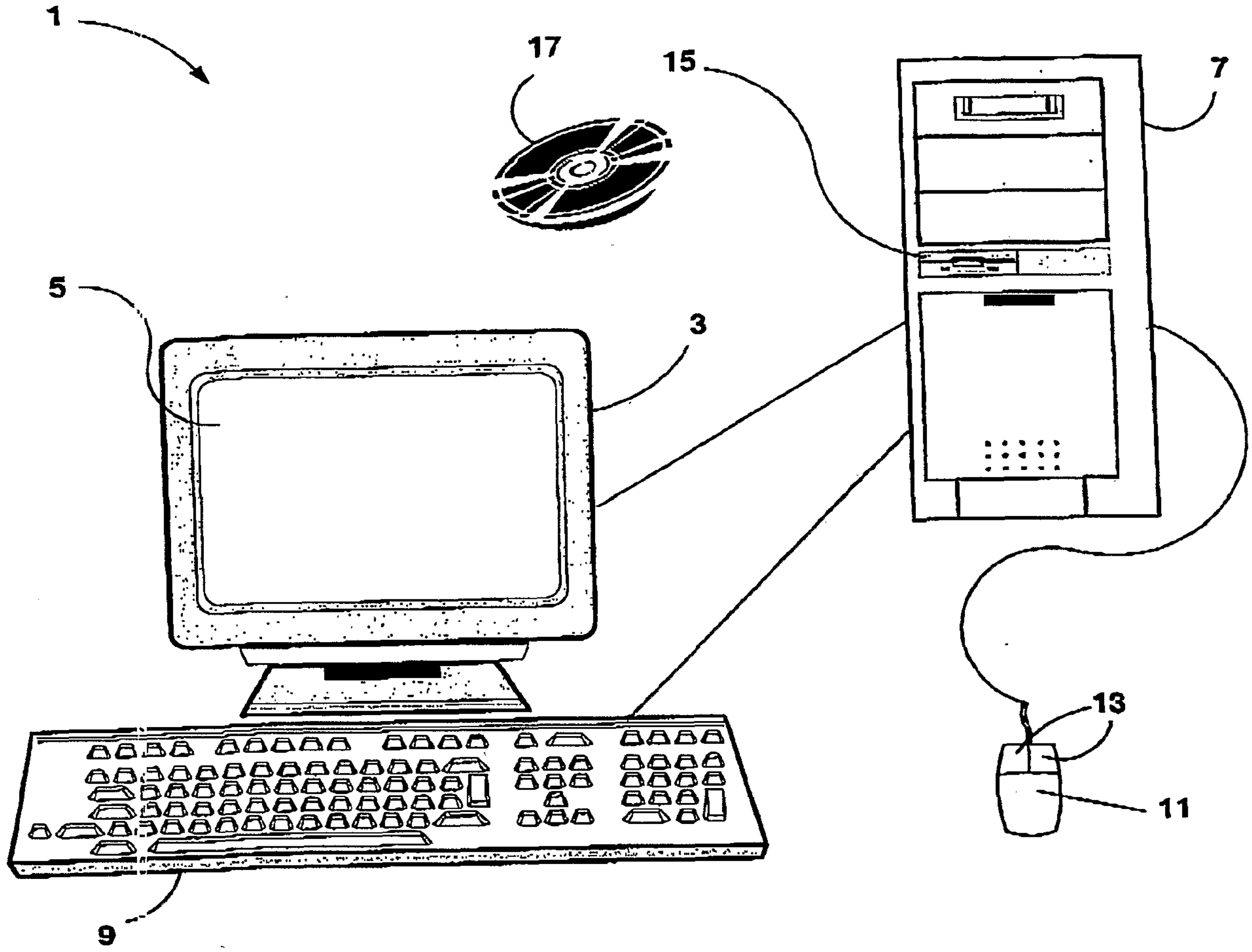


Figure 1

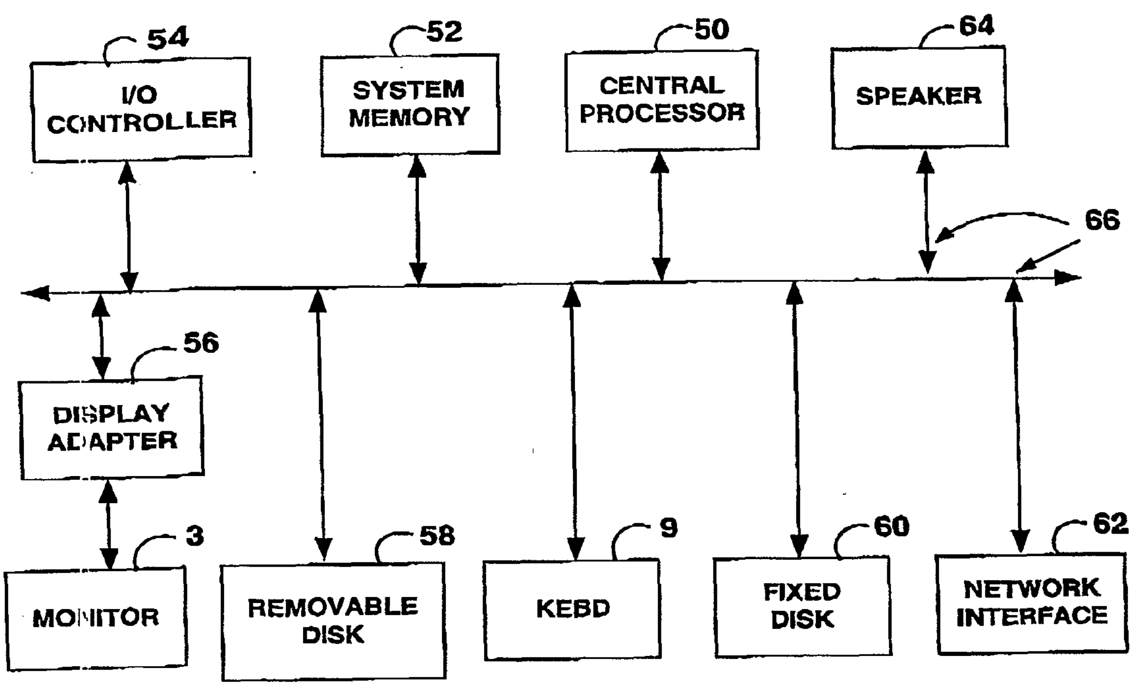
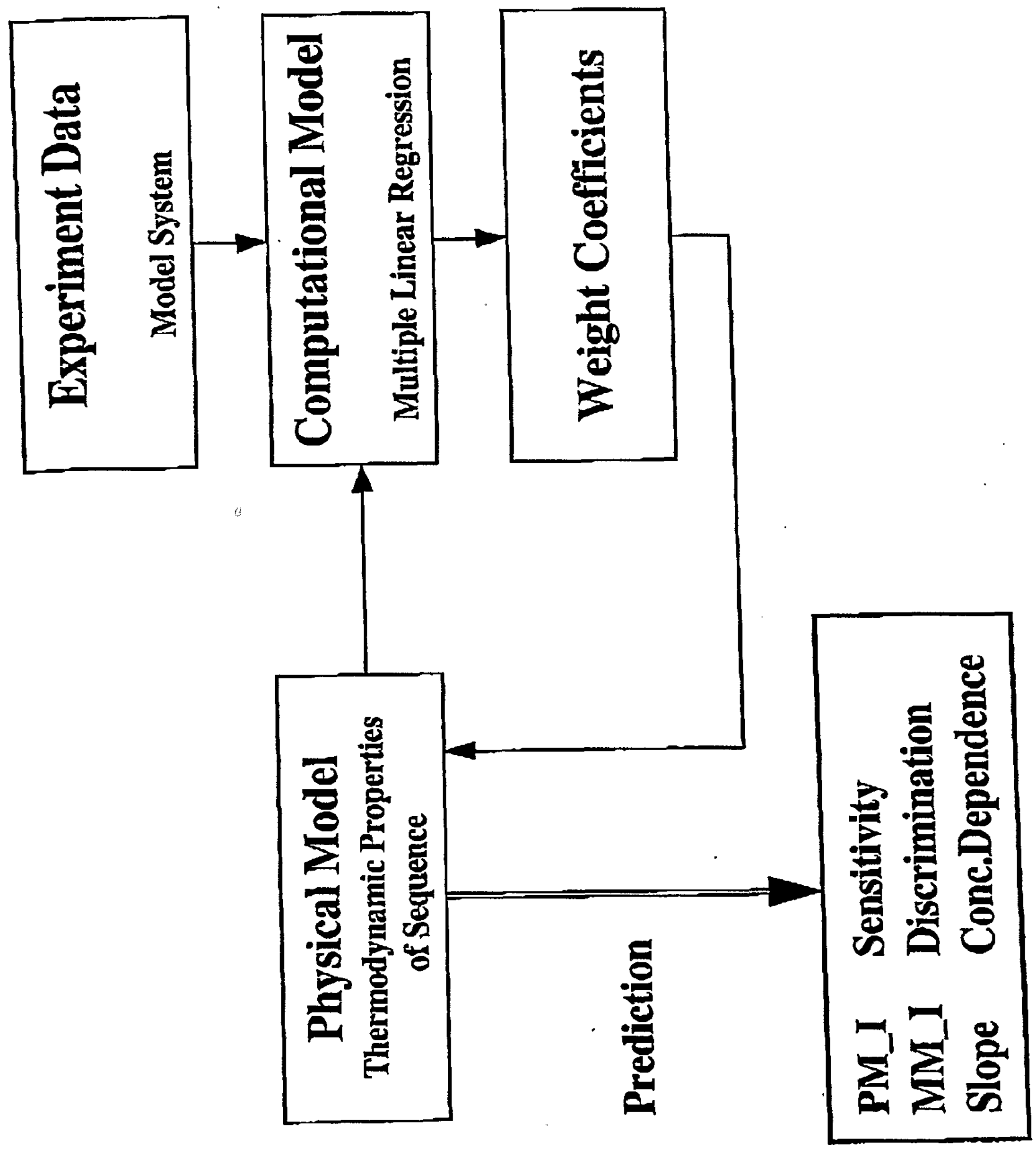

54
I/O CONTROLLER
52
SYSTEM MEMORY
50
CENTRAL PROCESSOR
64
SPEAKER
66
56
DISPLAY ADAPTER
3
MONITOR
58
REMOVABLE DISK
9
KEBD
60
FIXED DISK
62
NETWORK INTERFACE

Figure 2

Figure 3

Predicting Probe Quality



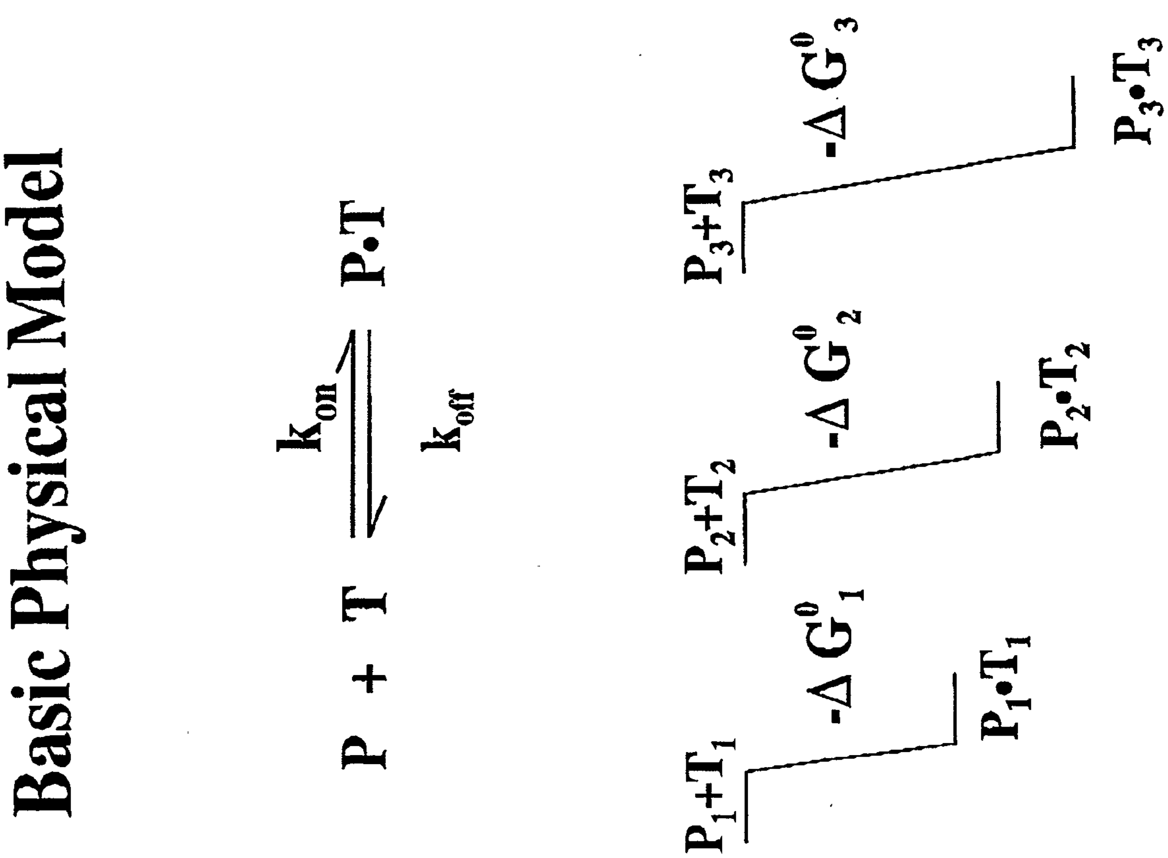

Basic Physical Model
P + T
k_on
k_off
P•T
P_1+T_1
-Δ G^0_1
P_1•T_1
P_2+T_2
-Δ G^0_2
P_2•T_2
P_3+T_3
-Δ G^0_3
P_3•T_3

Figure 4

Define Each Nucleotide at Each position

Figure 5

Example : GTCA

*Using A as ref. 3 base/position

i	Position	Base	S_i	
1	1	C	0	
2	1	G	1	(1st position is G)
3	1	T	0	
4	2	C	0	
5	2	G	0	
6	2	T	1	(2nd position is T)
7	3	C	1	(3rd position is C)
8	3	G	0	
9	3	T	0	
10	4	C	0	(4th position is A as reference)
11	4	G	0	
12	4	T	0	

Relative ΔG vs. Base Position

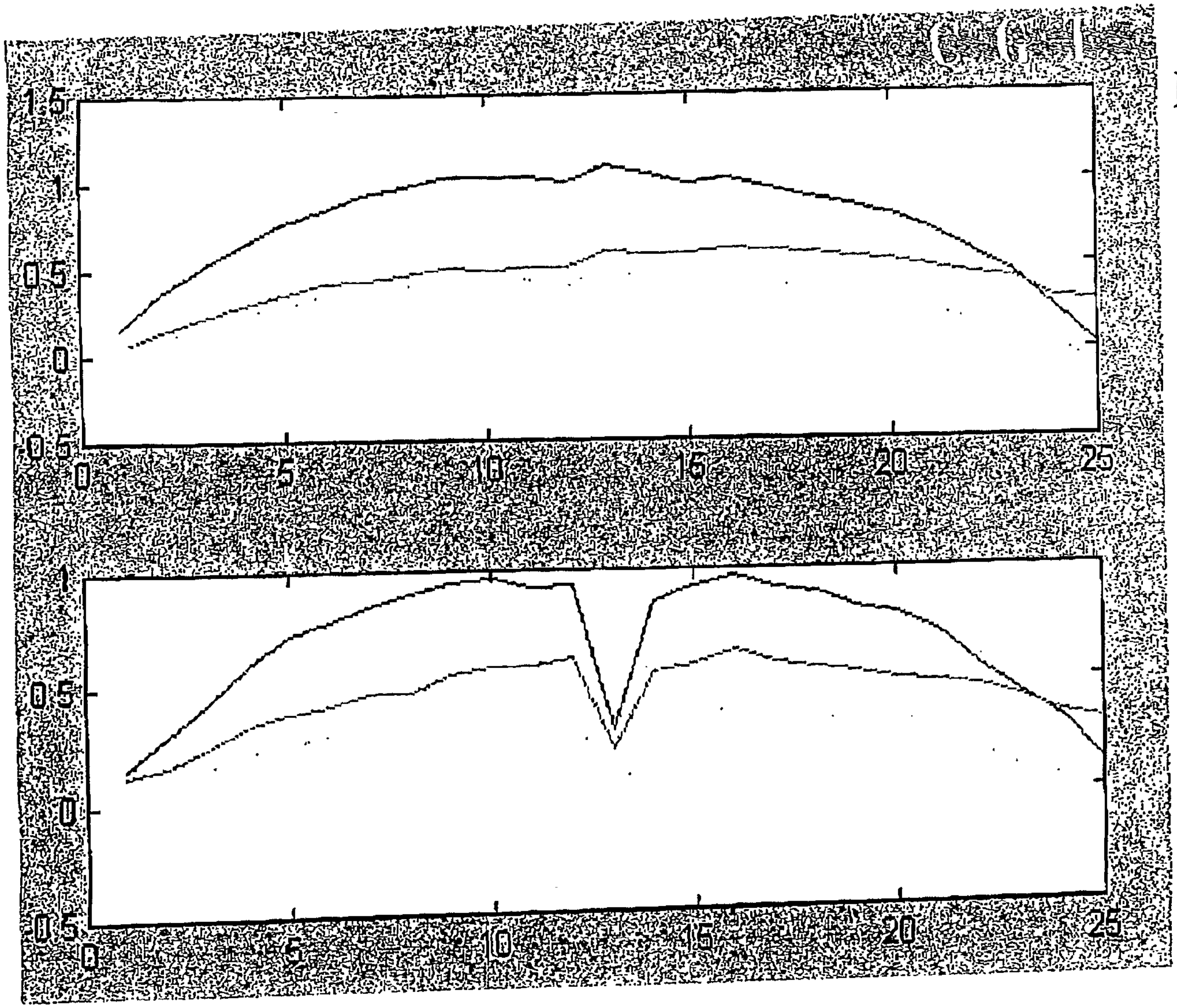


Figure 6A

PM

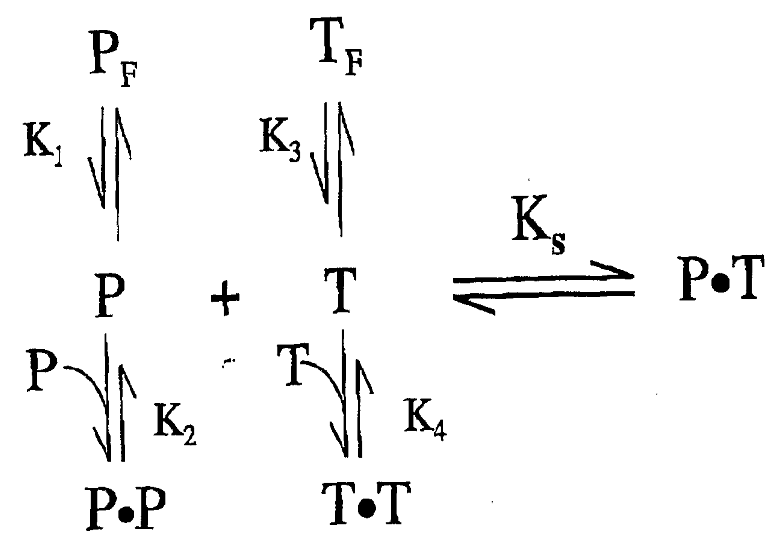
Figure 6B

MM

Base Position in Probe Sequence

Figure 7

Overall Reaction



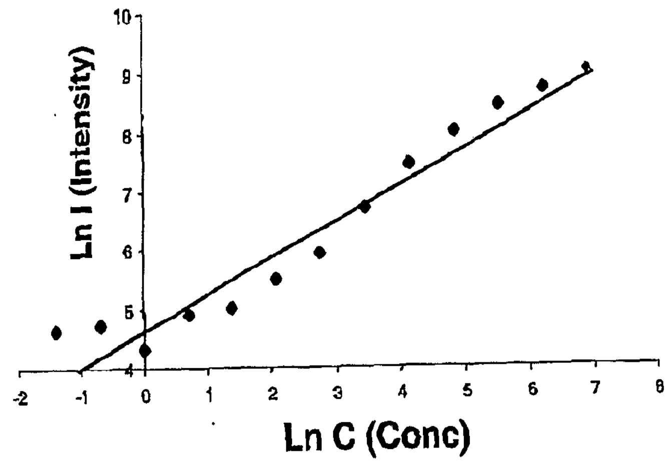
ΔG^0_{PF}: probe folding (K_1)
ΔG^0_{PP}: probe dimerization (K_2)
ΔG^0_{TF}: target folding (K_3)
ΔG^0_{TT}: target interaction (K_4)
ΔG^0_d: **duplex formation**

Concentration Dependence: Slope

Figure 8

$$\text{Ln I} = \text{S}*\text{Ln C} + \text{Ln K}_{app}$$

$$\text{I} = \text{K}_{app}*\text{C}^{S}$$



I: Intensity

K_{app}: Apparent Affinity Constant

C: Concentration

S: Empirical Value ($0 < S < 1$)

Relationship between Kapp vs. S

- Prediction of Probe Saturation

Figure 9A

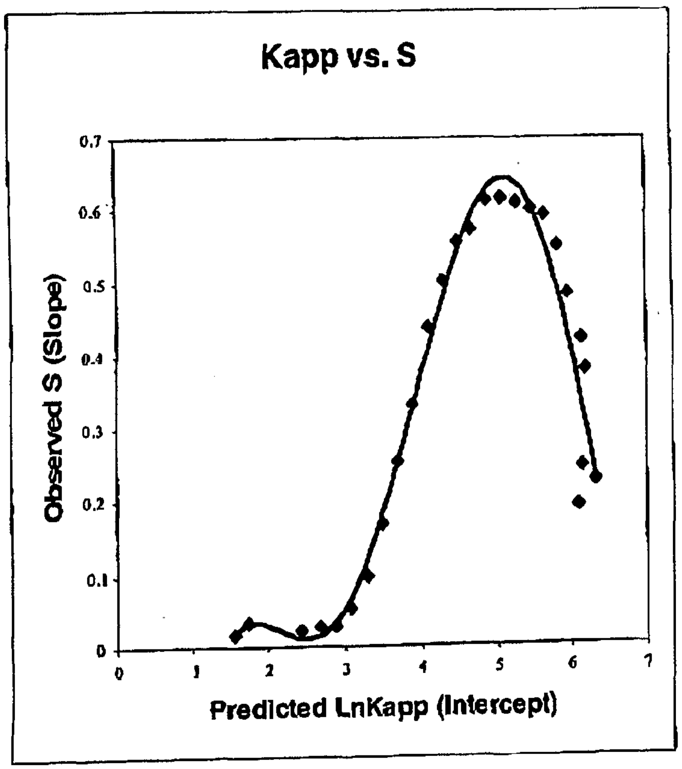
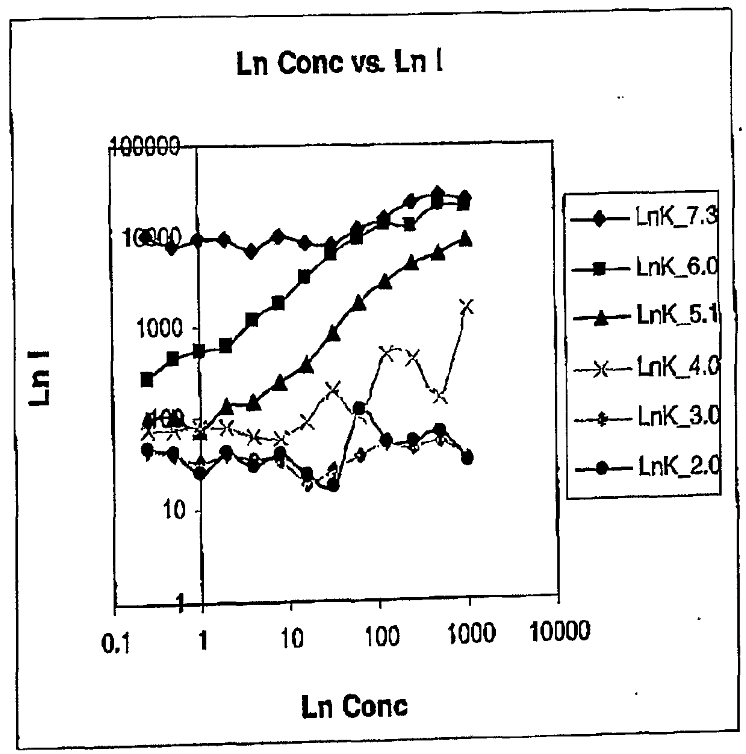


Figure 9B



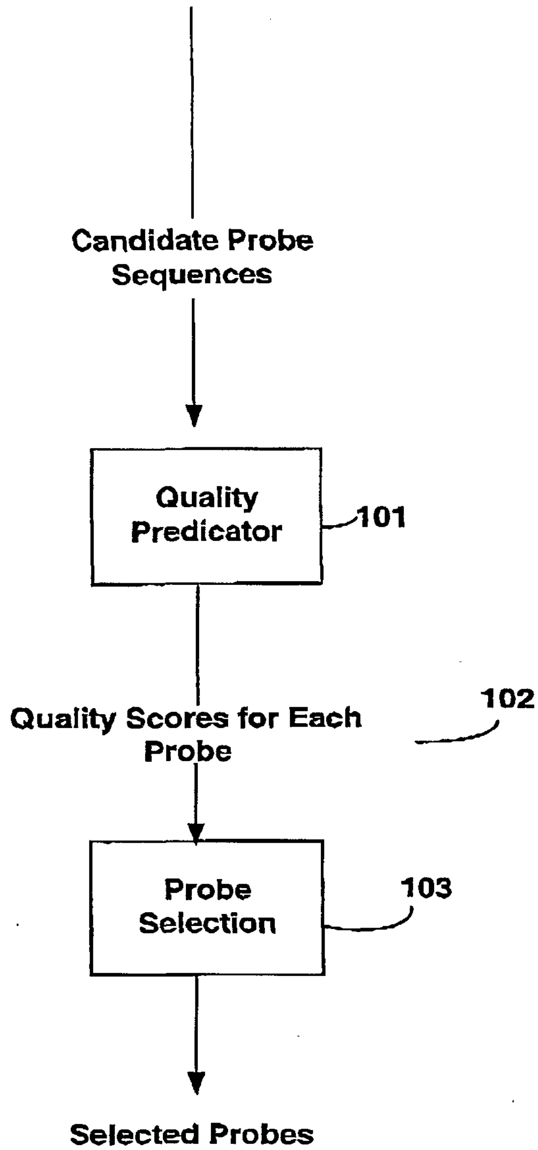

Candidate Probe Sequences
Quality Predicator
101
Quality Scores for Each Probe
102
Probe Selection
103
Selected Probes

Figure 10

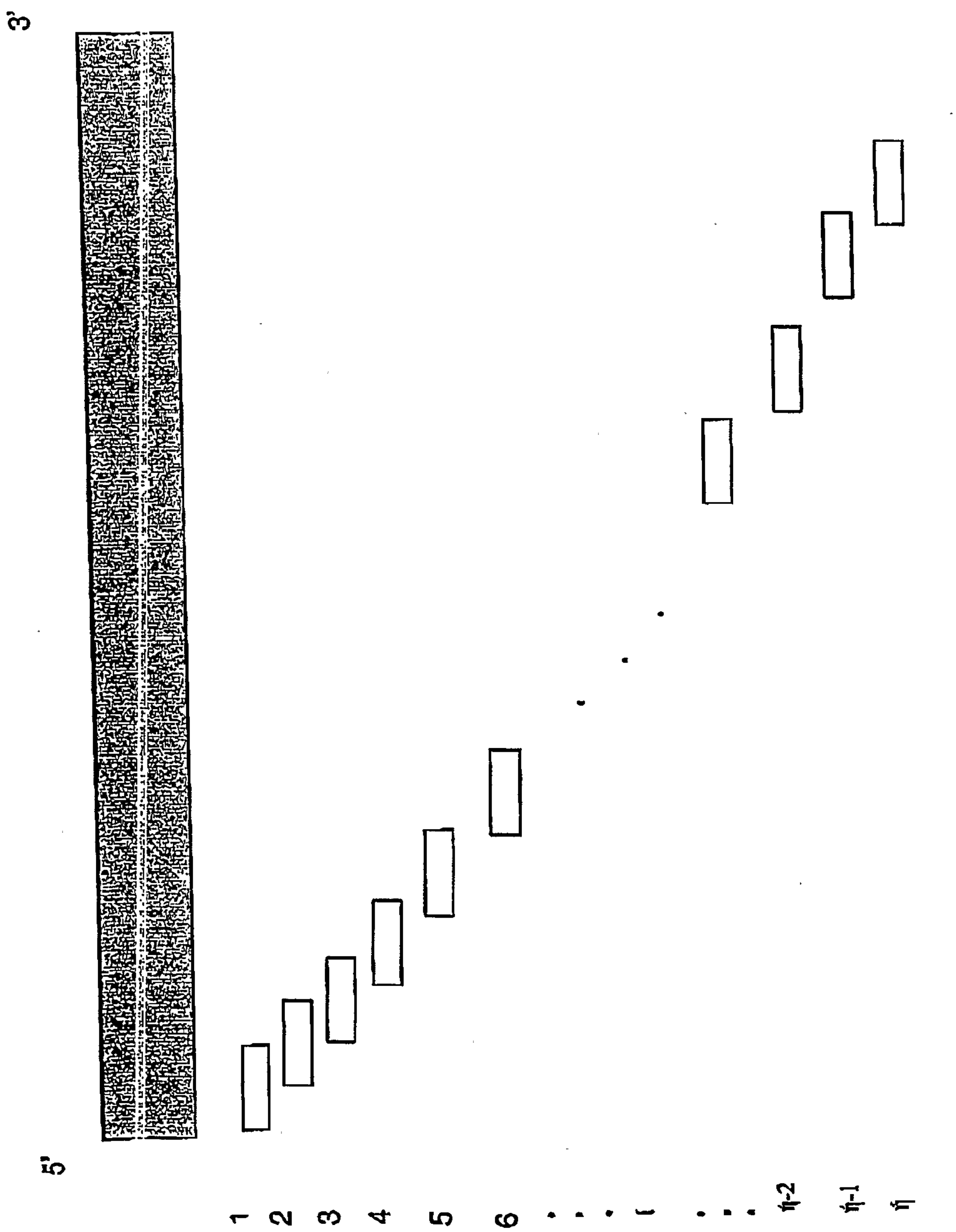


FIGURE 11

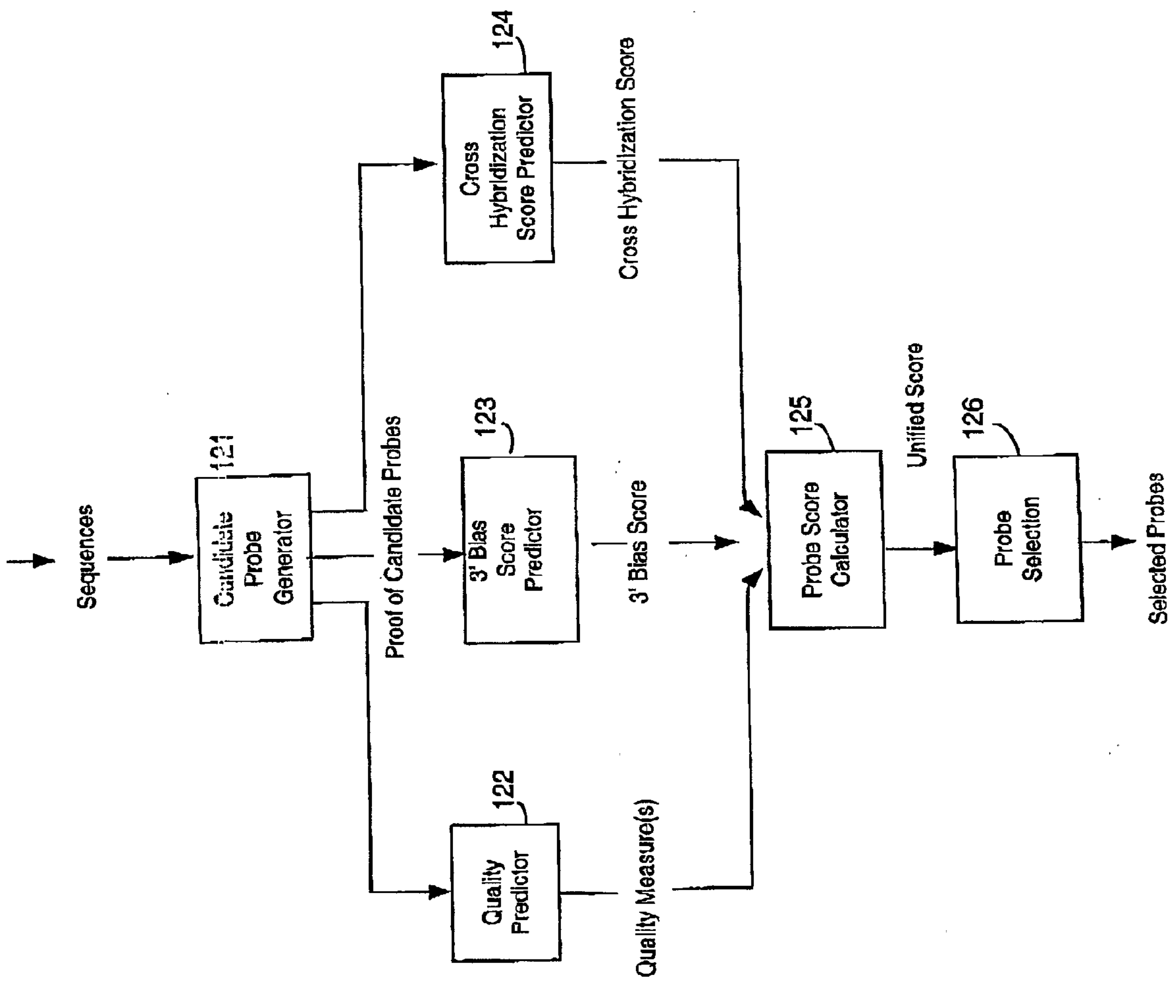

Sequences
Candidate Probe Generator
121
Proof of Candidate Probes
Quality Predictor
122
3' Bias Score Predictor
123
Cross Hybridization Score Predictor
124
Quality Measure(s)
3' Bias Score
Cross Hybridization Score
Probe Score Calculator
125
Unified Score
Probe Selection
126
Selected Probes

Figure 12

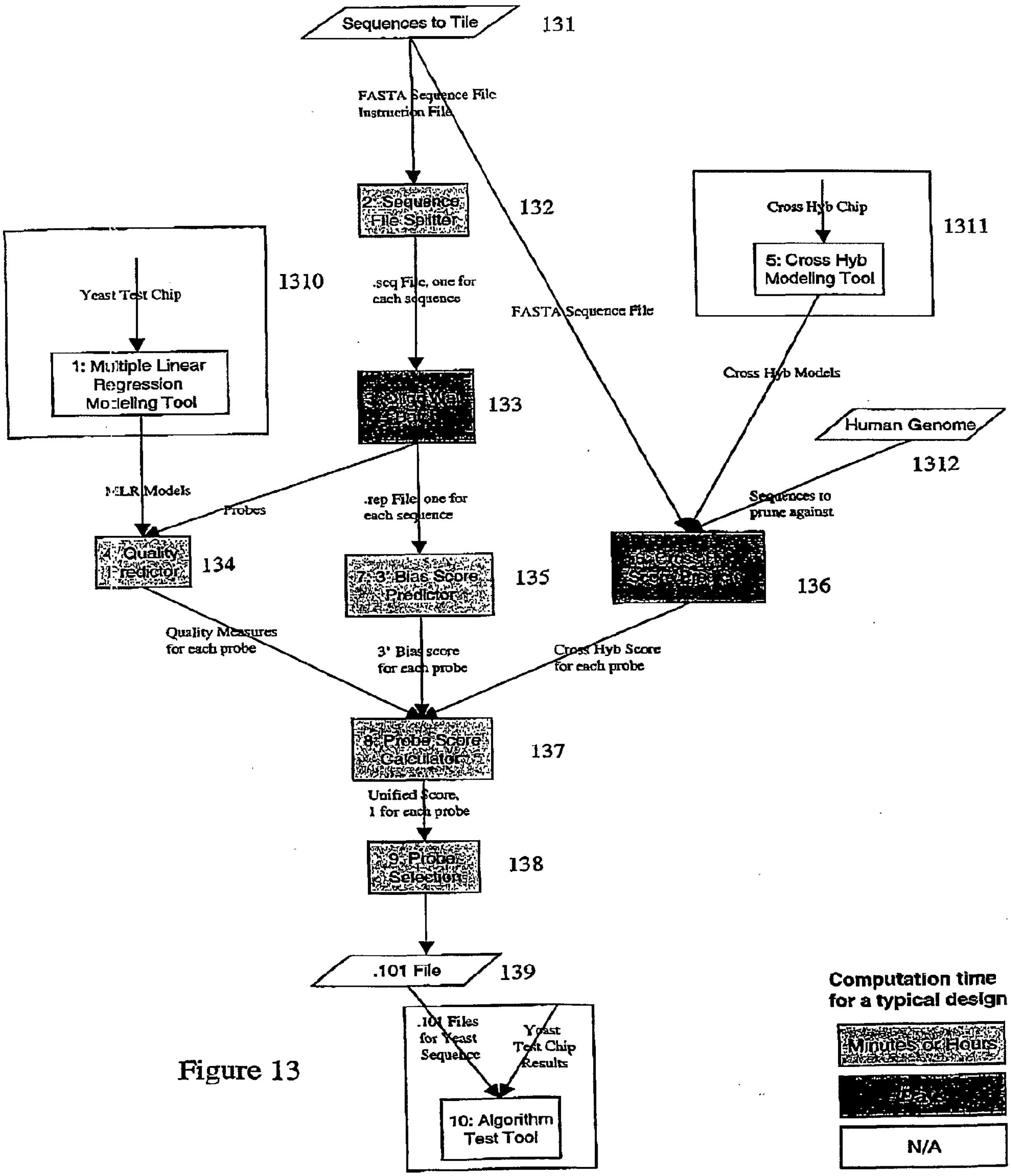

Sequences to Tile
131
FASTA Sequence File
Instruction File
2: Sequence File Splitter
132
.seq File, one for each sequence
FASTA Sequence File
Cross Hyb Chip
5: Cross Hyb Modeling Tool
1311
Yeast Test Chip
1: Multiple Linear Regression Modeling Tool
1310
133
Cross Hyb Models
Human Genome
1312
MLR Models
Probes
.rep File one for each sequence
Sequences to prune against
4: Quality Predictor
134
7: 3' Bias Score Predictor
135
136
Quality Measures for each probe
3' Bias score for each probe
Cross Hyb Score for each probe
8: Probe Score Calculator
137
Unified Score, 1 for each probe
9: Probe Selection
138
.101 File
139
.101 Files for Yeast Sequence
Yeast Test Chip Results
10: Algorithm Test Tool
Computation time for a typical design
Minutes or Hours
N/A

Figure 13

Latin Square MLR

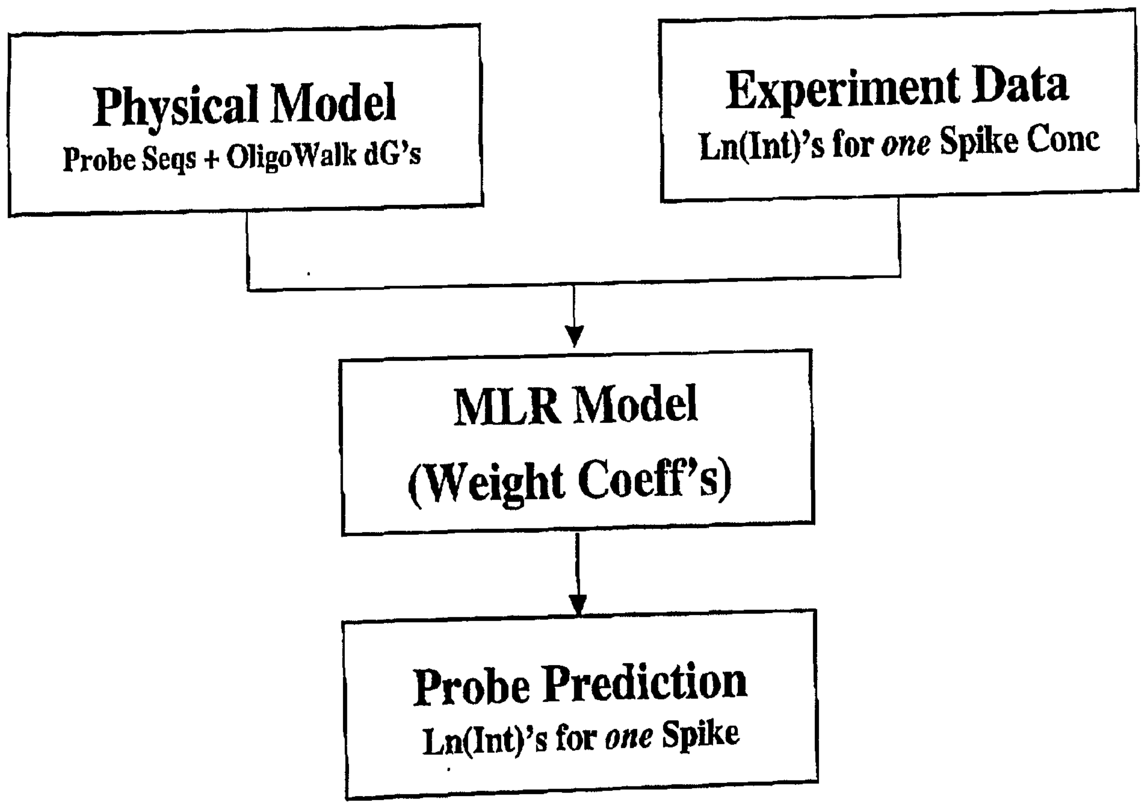


Figure 14

112 Yeast Clones Randomly Divided into 14 Groups

Groups

1	2	3	4	5	6	7	8	9	10	11	12	13	14
YNL259C	YNL037C	YAL038W	YHR044C	YMR127C	YLR377C	YOL064C	YPL209C	YIR034C	YJR148W	YEL046C	YGR185C	YBR166C	YOL165C
YEL003W	YDR113C	YLR083C	YJL117W	YNL290W	YOL086C	YJR094C	YFL029C	YMR276W	YML060W	YGR072W	YGL181W	YJL155C	YNL227C
YDL235C	YGL105W	YLL043W	YMR116C	YMR228W	YJR019C	YIR026C	YGR040W	YMR294W	YDL188C	YMR203W	YGL213C	YEL036C	YNL228W
YEL024W	YDR498C	YBR212W	YPL111W	YPR057W	YOR085W	YLR056W	YPR065W	YPL001W	YGR109C	YGR112W	YOL136C	YJL014W	YMR108W
YEL018W	YDL029W	YNL015W	YCL055W	YNR035C	YDL226C	YMR270C	YPR191W	YEL039C	YOL043C	YHR208W	YEL037C	YJL110C	YPL043W
YER161C	YKL081W	YDL075W	YFR025C	YCL032W	YBL016W	YBR018C	YMR139W	YNL307C	YLR291C	YIL136W	YHL022C	YFL056C	YLR153C
YKL193C	YFR053C	YML098W	YLR354C	YIL154C	YBL068W	YBR057C	YPR035W	YGL148W	YDR086C	YOR099W	YHL014C	YJR155W	YPR074C
YPR129W	YFL018C	YOL143C	YPL069C	YBR034C	YHR025W	YER118C	YNL005C	YGL155W	YNR015W	YOR176W	YKR061W	YNL331C	YPL089C

Figure 15

Latin Square Experiment

Groups →

Exp ↓

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	0.25	0.5	1	2	4	8	16	32	64	128	256	512	1024
2	0.25	0.5	1	2	4	8	16	32	64	128	256	512	1024	0
3	0.5	1	2	4	8	16	32	64	128	256	512	1024	0	0.25
4	1	2	4	8	16	32	64	128	256	512	1024	0	0.25	0.5
5	2	4	8	16	32	64	128	256	512	1024	0	0.25	0.5	1
6	4	8	16	32	64	128	256	512	1024	0	0.25	0.5	1	2
7	8	16	32	64	128	256	512	1024	0	0.25	0.5	1	2	4
8	16	32	64	128	256	512	1024	0	0.25	0.5	1	2	4	8
9	32	64	128	256	512	1024	0	0.25	0.5	1	2	4	8	16
10	64	128	256	512	1024	0	0.25	0.5	1	2	4	8	16	32
11	128	256	512	1024	0	0.25	0.5	1	2	4	8	16	32	64
12	256	512	1024	0	0.25	0.5	1	2	4	8	16	32	64	128
13	512	1024	0	0.25	0.5	1	2	4	8	16	32	64	128	256
14	1024	0	0.25	0.5	1	2	4	8	16	32	64	128	256	512

Figure 16

Latin Square Data Sets from Yeast_Test_Hyb Chips

Lot 1 (9912072)

No Background:	3 Scans	14 chips	(530, PMT=701; 570, PMT=701; 570, PMT=600)
+ Background:	3 Scans	14 chips	(530, PMT=701; 570, PMT=701; 570, PMT=600)

Lot 2 (9910426)

No Background:	1 Scan	14 chips	(570, PMT=600)
No Background:	1 Scan	14 chips	(570, PMT=600)

Lot 3 (9910427)

+ Background:	1 Scan	14 chips	(570, PMT=526)
No Background_Rep1:	1 Scan	14 chips	(570, PMT=526)
No Background_Rep2:	1 Scan	14 chips	(570, PMT=526)

Lot 4 (9913514)

No Background:	1 Scan	14 chips	(570, PMT=526)
+ Background:	1 Scan	14 chips	(570, PMT=526)

Lot 5 (9914059)

+ Background_Rep1:	1 Scan	14 chips	(570, PMT=526)
+ Background_Rep2:	1 Scan	14 chips	(570, PMT=526)
No Background:	1 Scan	14 chips	(570, PMT=526)

Figure 17

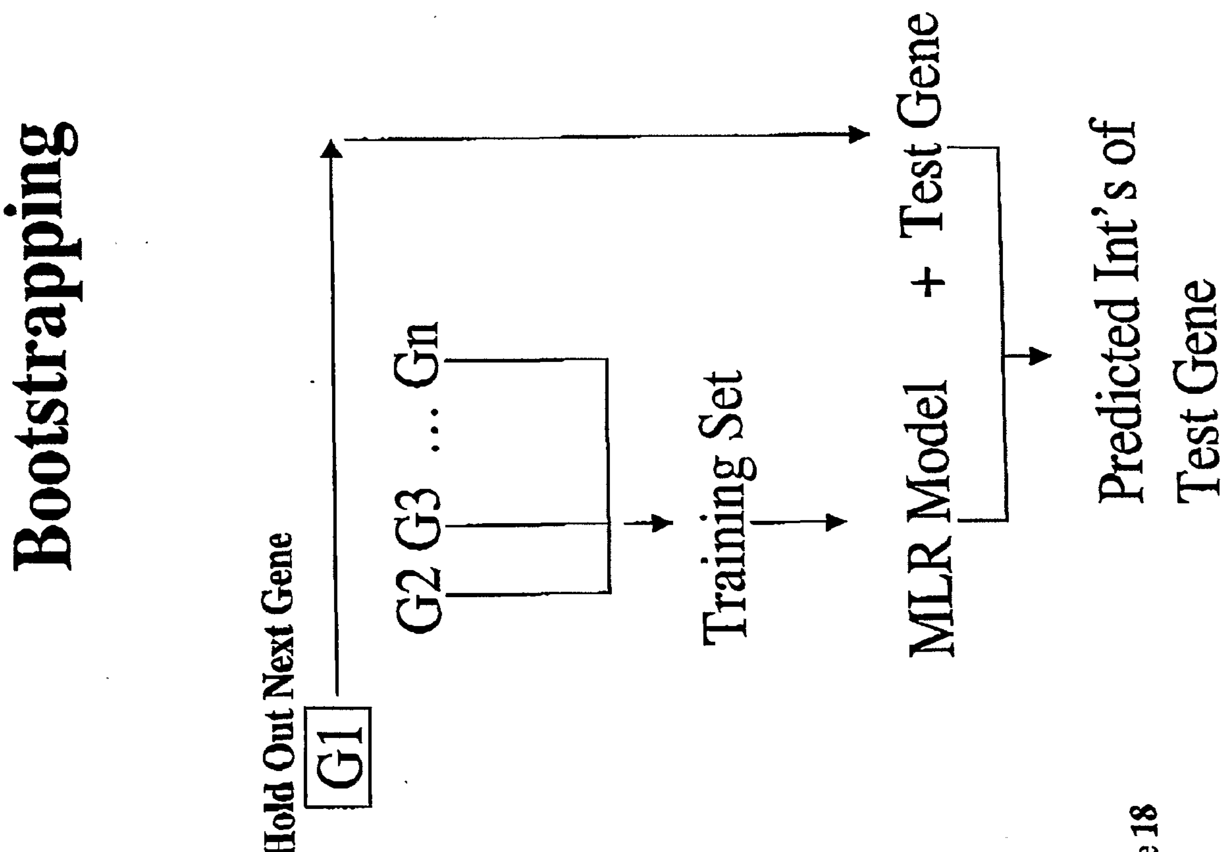

Bootstrapping
Hold Out Next Gene
G1
G2 G3 ... Gn
Training Set
MLR Model
+ Test Gene
Predicted Int's of
Test Gene

Figure 18

YDR113C

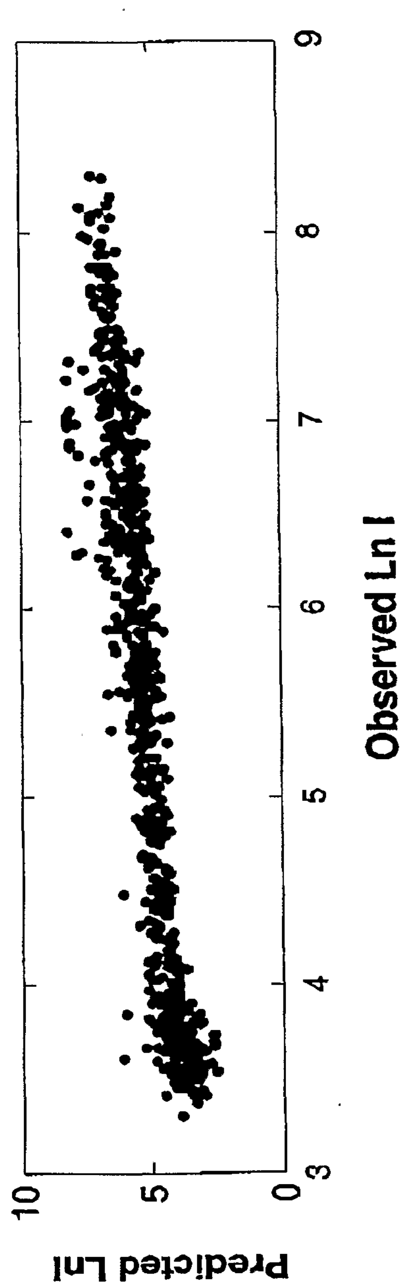


Figure 19A

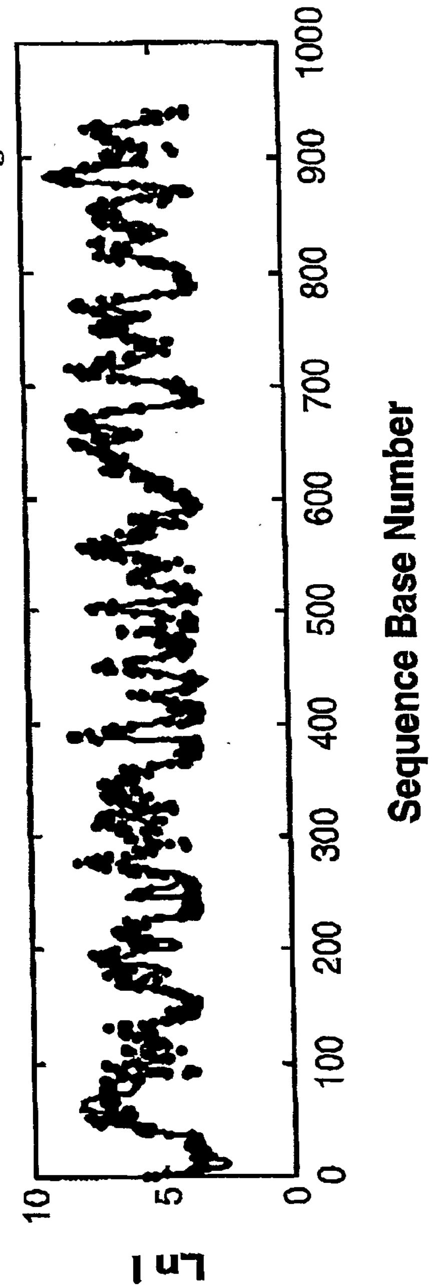


Figure 19B

Figure 20A

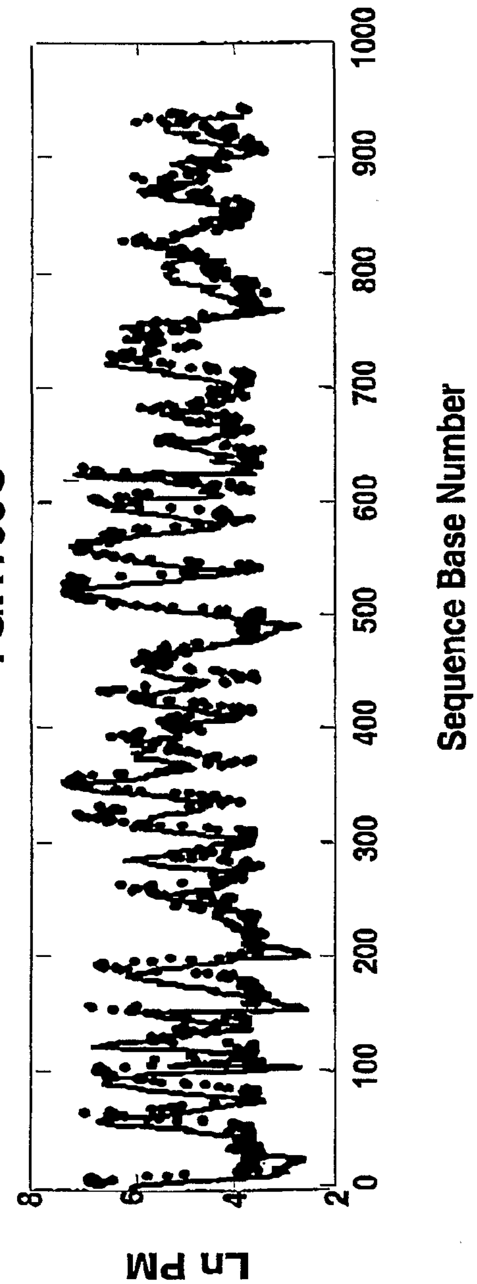
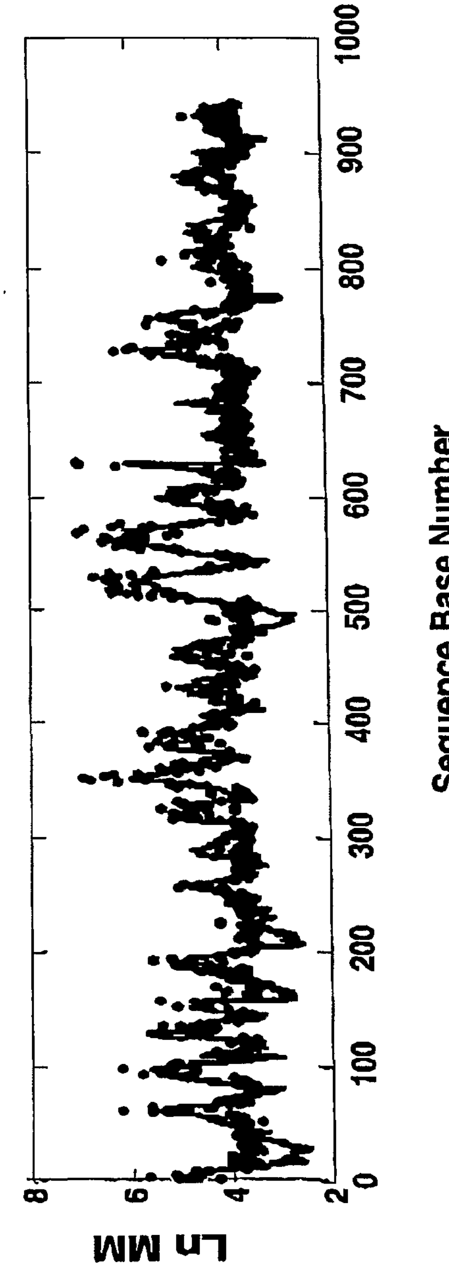


Figure 20B



Ln(Int) at Different Spike Concentrations

8pM

16pM

32pM

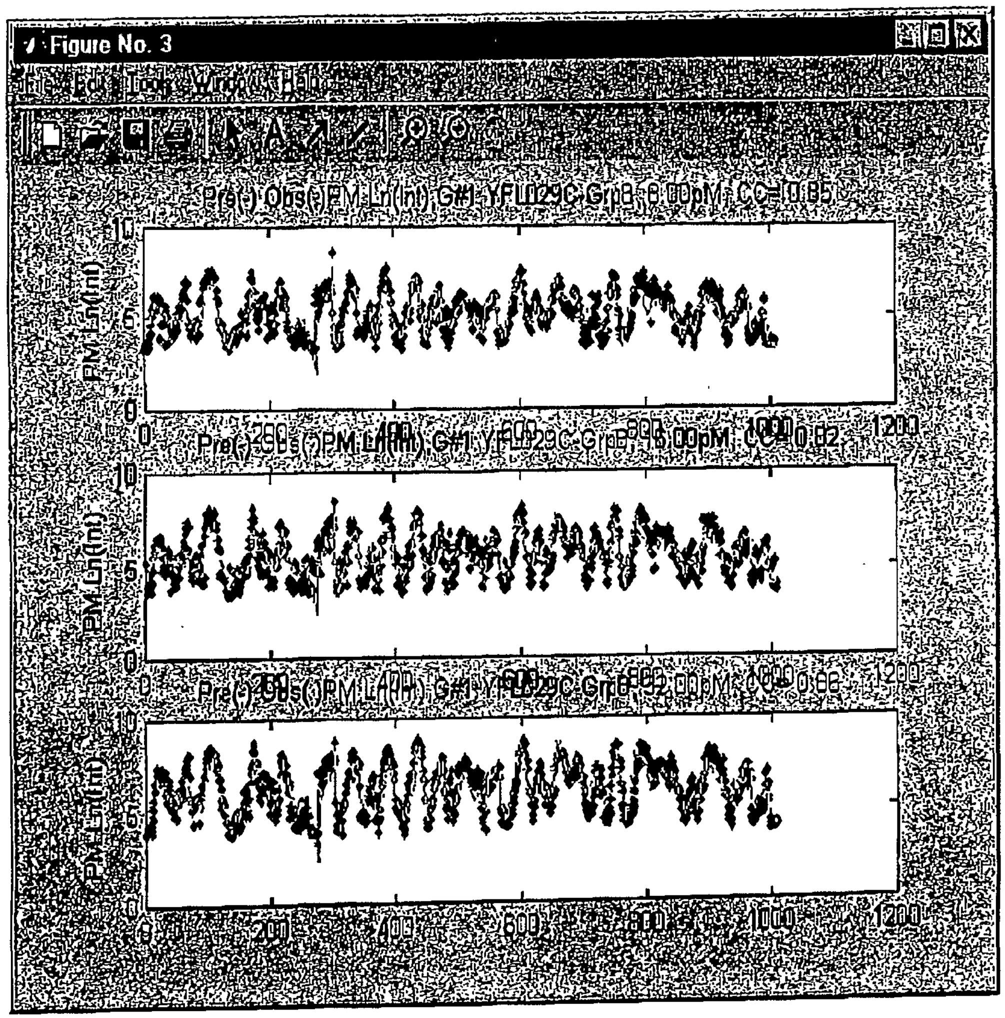


Figure 21

Correlation between Predicted & Observed Ln(Int)'s

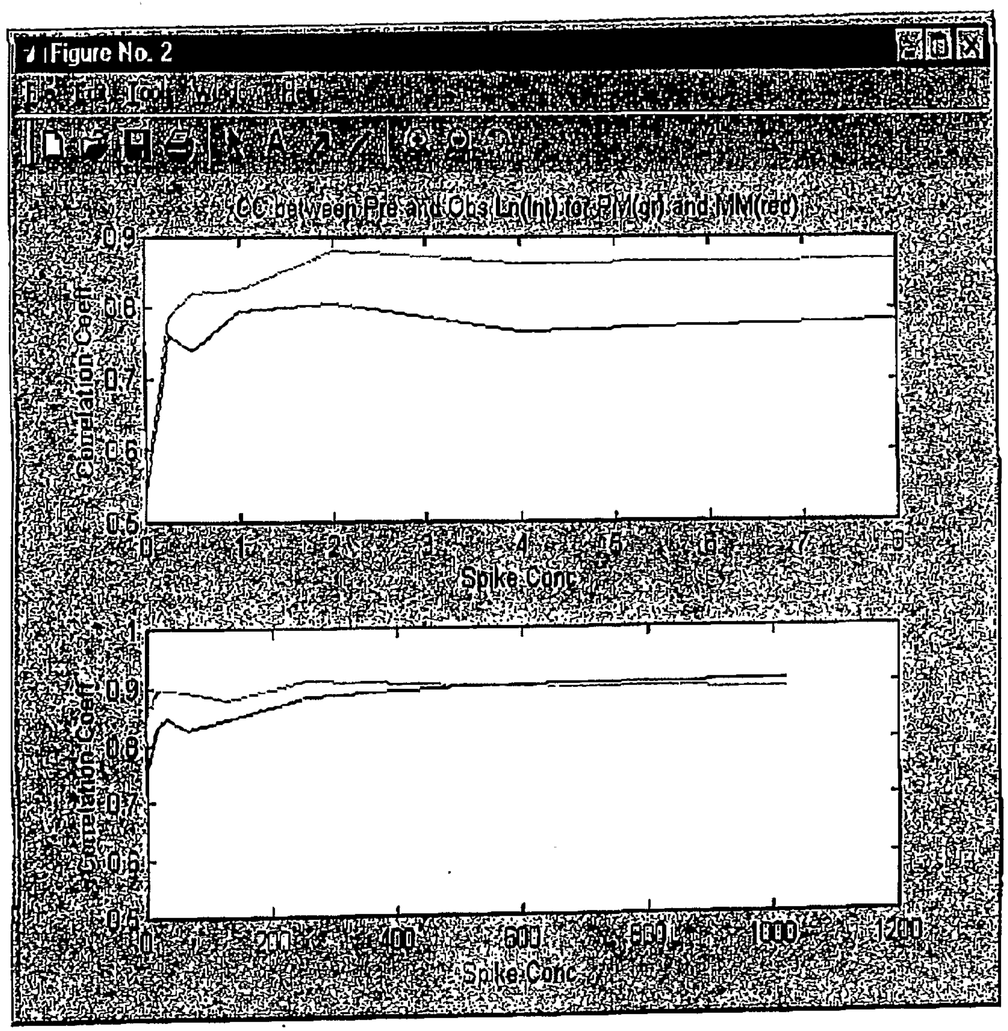


Figure 22

Negative Control: Gene in Wrong Orientation

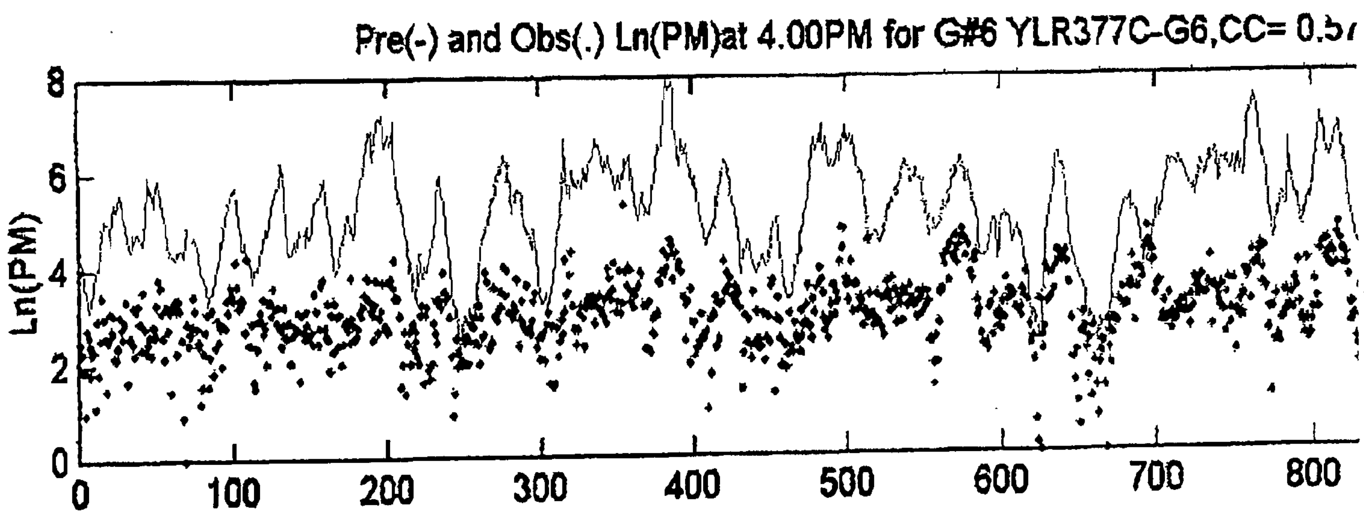


Figure 23

Predicted Observed Slopes

CC=0.42

CC=0.84

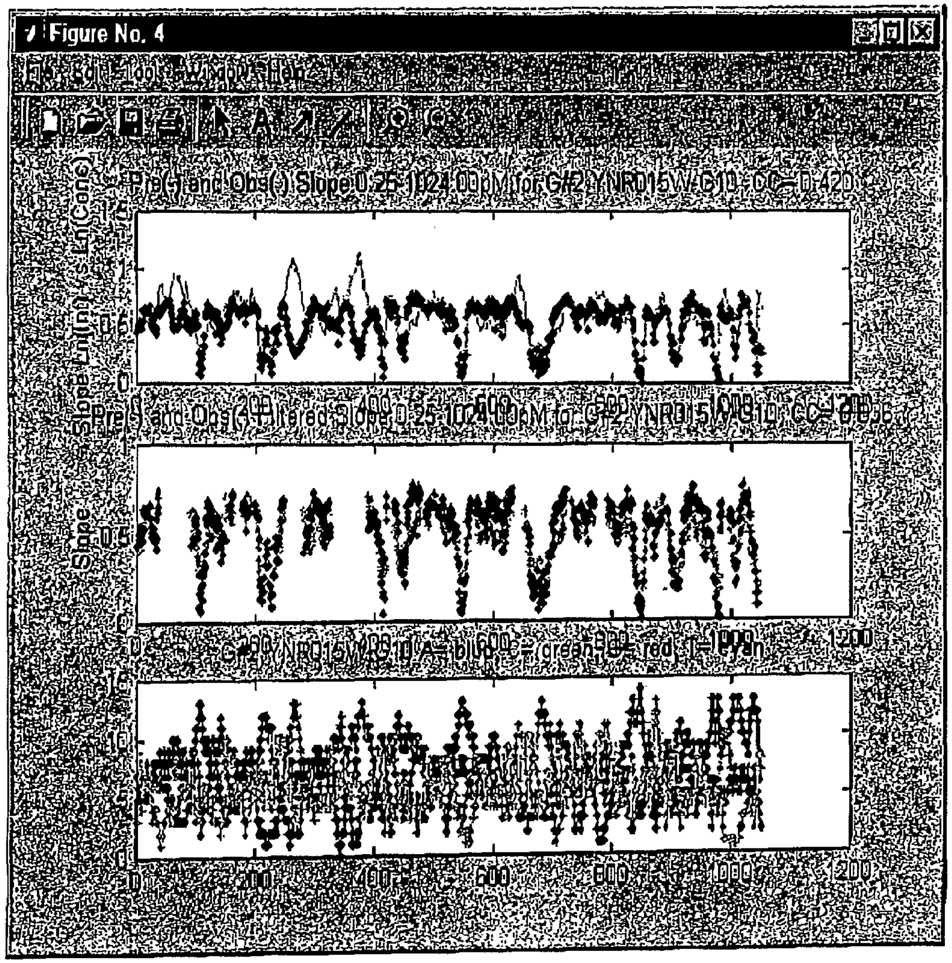
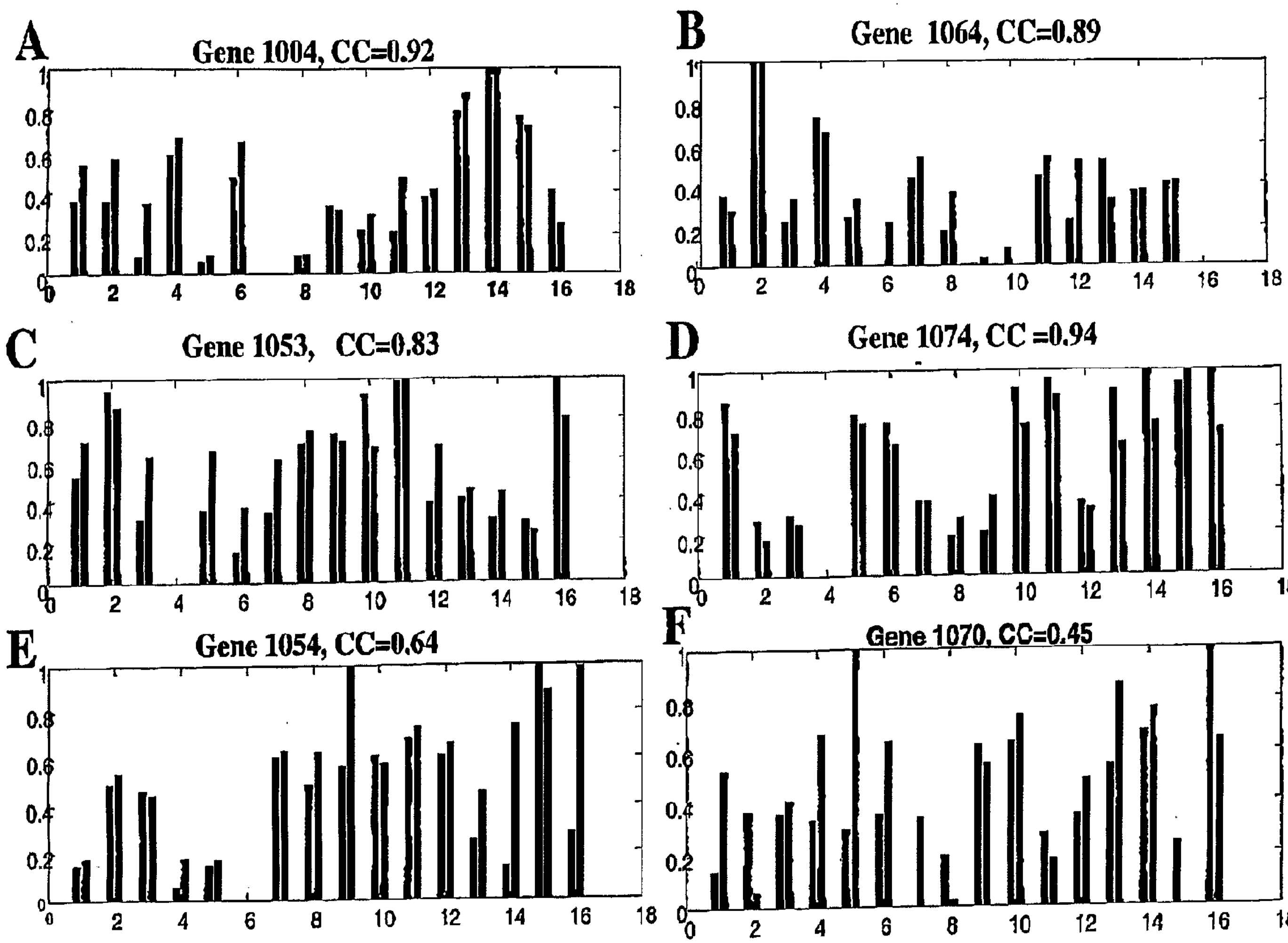


Figure 24

From Yeast to Human

Figure 25

Using Parameters from Yeast Model System to Predict Human_U95A



Predictions for Hu_U95a Probe Sets

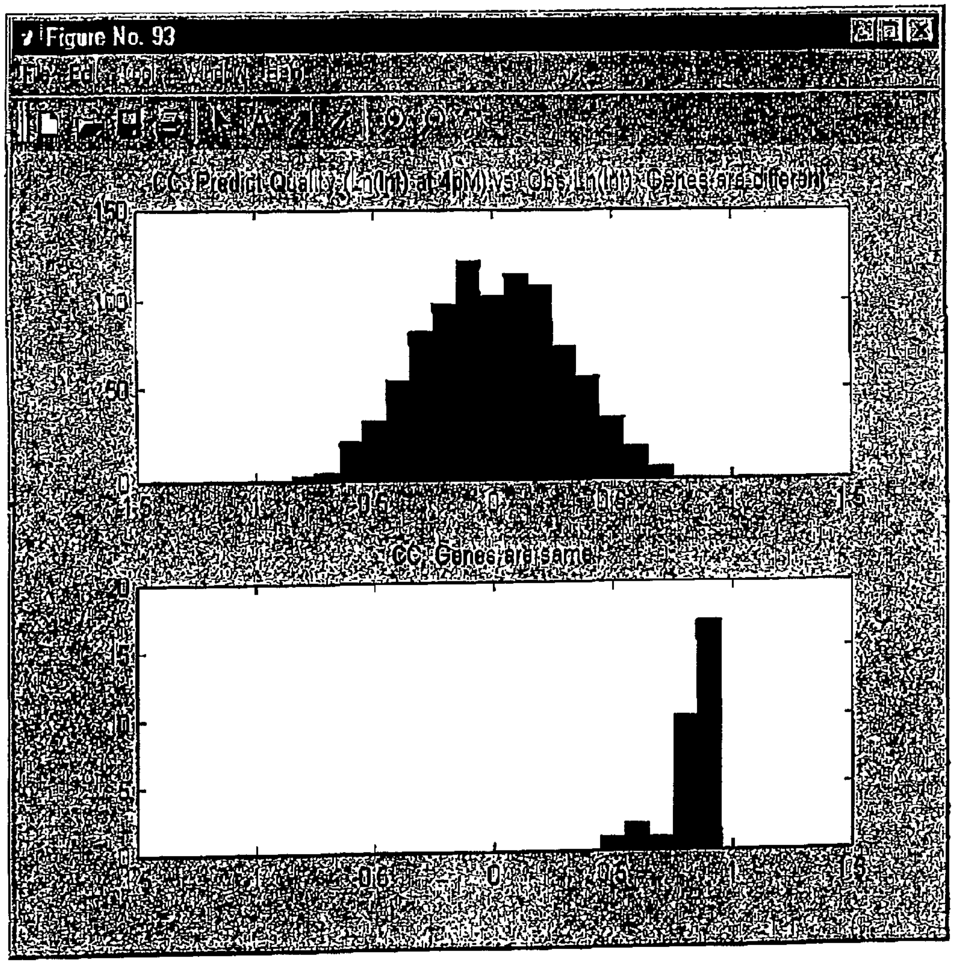


Figure 26

Sixteen Probes Selected By Dynamic Programming Algorithm

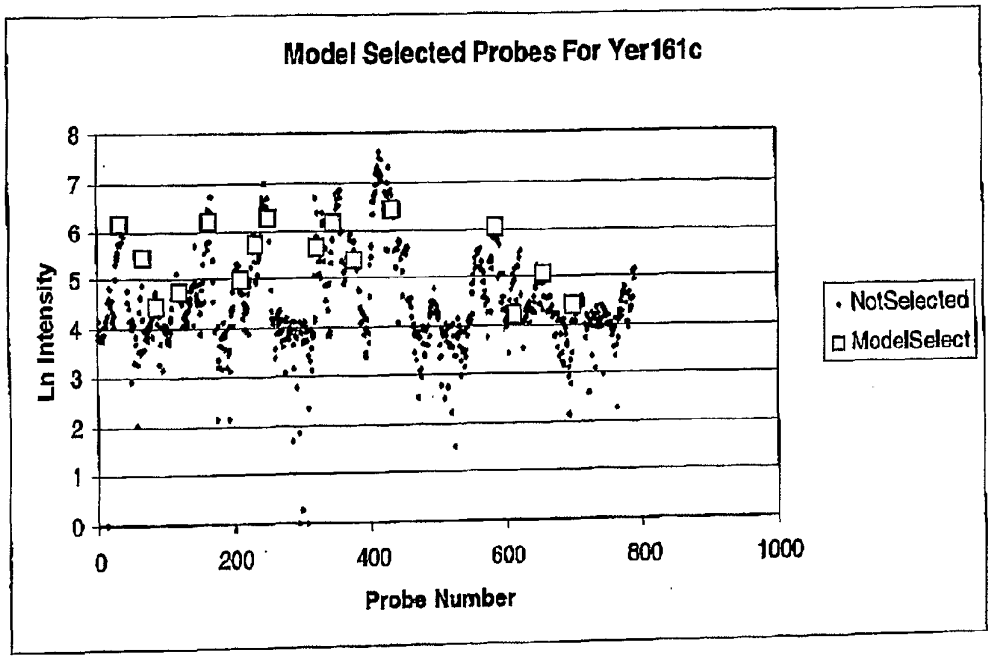


Figure 27

Comparison of AveDiff Values of all Yeast Test Chip Genes:New vs Random vs Rules Selection

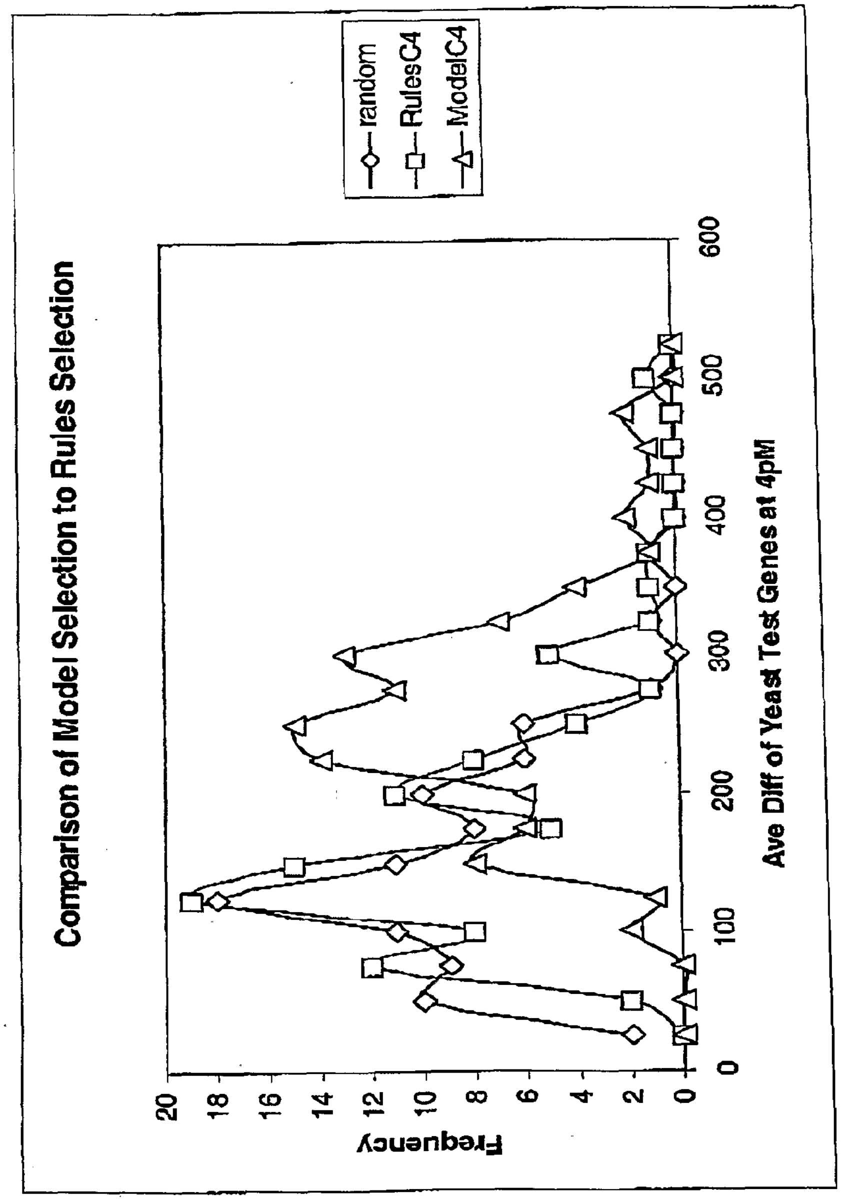


Figure 28